1 **Realised ecological forecast through interactive Ecological Platform for Assimilating Data**

2 **into model (EcoPAD (v1.0))**

3

4 Yuanyuan Huang[1,2], Mark Stacy[3], Jiang Jiang[1,4], Nilutpal Sundi[5], Shuang Ma[1,6], Volodymyr

5 Saruta[1,6], Chang Gyo Jung[1,6], Zheng Shi[1], Jianyang Xia[7,8], Paul J.Hanson[9], Daniel Ricciuto[9], Yiqi

6 Luo[1,6,10]

7

8 1, Department of Microbiology and Plant Biology, University of Oklahoma, Norman, Oklahoma

9 2, Laboratoire des Sciences du Climat et de l'Environnement, 91191 Gif-sur-Yvette, France

10 3, University of Oklahoma Information Technology, Norman, Oklahoma, USA

11 4, Key Laboratory of Soil and Water Conservation and Ecological Restoration in Jiangsu Province,

12 Collaborative Innovation Center of Sustainable Forestry in Southern China of Jiangsu Province, Nanjing

13 Forestry University, Nanjing, Jiangsu, China

14 5, Department of Computer Science, University of Oklahoma, Norman, Oklahoma, USA

15 6, Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, AZ, USA

16 7, Tiantong National Forest Ecosystem Observation and Research Station, School of Ecological and

17 Environmental Sciences, East China Normal University, Shanghai 200062, China.

18 8, Research Center for Global Change and Ecological Forecasting, East China Normal

19 University, Shanghai 200062, China

20 9, Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National

21 Laboratory, Oak Ridge, Tennessee, USA

22 10, Department of Earth System Science, Tsinghua University, Beijing 100084 China

23

24 Correspondence: Yuanyuan Huang (yuanyuanhuang2011@gmail.com) and Yiqi Luo
25 (Yiqi.Luo@nau.edu)

26   **Abstract.** Predicting future changes in ecosystem services is not only highly desirable but also

27   becomes feasible as several forces (e.g., available big data, developed data assimilation (DA)

28   techniques, and advanced cyberinfrastructure) are converging to transform ecological research to

29   quantitative forecasting. To realise ecological forecasting, we have developed an Ecological

30   Platform for Assimilating Data (EcoPAD (v1.0)) into models. EcoPAD (v1.0) is a web-based

31   software system that automates data transfer and processing from sensor networks to ecological

32   forecasting through data management, model simulation, data assimilation, forecasting and

33   visualization. It facilitates interactive data-model integration from which model is recursively

34   improved through updated data while data is systematically refined under the guidance of model.

35   EcoPAD (v1.0) relies on data from observations, process-oriented models, DA techniques, and

36   the web-based workflow.

37   We applied EcoPAD (v1.0) to the Spruce and Peatland Responses Under Climatic and

38   Environmental change (SPRUCE) experiment at North Minnesota. The EcoPAD-SPRUCE

39   realises fully automated data transfer, feeds meteorological data to drive model simulations,

40   assimilates both manually measured and automated sensor data into Terrestrial ECOsystem

41   (TECO) model, and recursively forecast responses of various biophysical and biogeochemical

42   processes to five temperature and two CO_2 treatments in near real-time (weekly). Forecasting

43   with EcoPAD-SPRUCE has revealed that mismatches in forecasting carbon pool dynamics are

44   more related to model (e.g., model structure, parameter, and initial value) than forcing variables,

45   opposite to forecasting flux variables. EcoPAD-SPRUCE quantified acclimations of methane

46   production in response to warming treatments through shifted posterior distributions of the

47   $CH_4:CO_2$ ratio and temperature sensitivity (Q_{10}) of methane production towards lower values.

48   Different case studies indicated that realistic forecasting of carbon dynamics relies on

49   appropriate model structure, correct parameterization and accurate external forcing. Moreover,

50   EcoPAD-SPRUCE stimulated active feedbacks between experimenters and modellers to identify

51   model components to be improved and additional measurements to be made. It becomes the

52   interactive model-experiment (ModEx) system and opens a novel avenue for interactive dialogue

53   between modellers and experimenters. Altogether, EcoPAD (v1.0) acts to integrate multiple

54   sources of information and knowledge to best inform ecological forecasting.

55

56

57   **Key words:**

58   Data assimilation, SPRUCE, carbon, global change, real time, acclimation, forecast

59

1. Introduction

One ambitious goal of ecology as a science discipline is to forecast states and services of ecological systems. Forecasting in ecology is not only desirable for scientific advances in this discipline but also has practical values to guide resource management and decision-making towards a sustainable planet Earth. The practical need for ecological forecasting is particularly urgent in this rapidly changing world, which is experiencing unprecedented natural resource depletion, increasing food demand, serious biodiversity crisis, accelerated climate changes, and widespread pollutions in the air, waters, and soils (Clark et al., 2001;Mouquet et al., 2015). As a result, a growing number of studies have reported forecasting of, e.g., phenology (Diez et al., 2012), carbon dynamics (Luo et al., 2016;Gao et al., 2011;Thomas et al., 2017), species dynamics (Clark et al., 2003;Kearney et al., 2010), pollinator performance (Corbet et al., 1995), epidemics (Ong et al., 2010), fishery (Hare et al., 2010), algal bloom (Stumpf et al., 2009), crop yield (Bastiaanssen and Ali, 2003), biodiversity (Botkin et al., 2007), plant extinction risk (Fordham et al., 2012), and ecosystem service (Craft et al., 2009) in the last several decades. Despite its broad applications, ecological forecasting is still sporadically practiced and lags far behind demand due to the lack of infrastructure that enables timely integration of models with data. This paper introduces the fully interactive infrastructure, the Ecological Platform for Assimilating Data (EcoPAD (v1.0)) into models, to inform near-time ecological forecasting with iterative data-model integration.

Ecological forecasting relies on both models and data. However, currently the ecology research community has not yet adequately integrated observations with models to inform best forecast. Forecasts generated from scenario approaches are qualitative and scenarios are often not based on ecological knowledge (Coreau et al., 2009;Coreau et al., 2010). Data-driven

83 forecasts using statistical methods are generally limited for extrapolation and sometimes

84 contaminated by confounding factors (Schindler and Hilborn, 2015). Recent emergent

85 mechanism-free non-parametric approach, which depends on the statistical pattern extracted

86 from data, is reported to be promising for short-term forecast (Ward et al., 2014;Perretti et al.,

87 2013;Sugihara et al., 2012), but has limited capability in long-term prediction due to the lack of

88 relevant ecological mechanisms. Process-based models provide the capacity in long-term

89 prediction and the flexibility in capturing short-term dynamics on the basis of mechanistic

90 understanding (Coreau et al., 2009;Purves et al., 2013). Wide applications of process-based

91 models are limited by their often complicated numerical structure and sometimes unrealistic

92 parameterization (Moorcroft, 2006). The complex and uncertain nature of ecology precludes

93 practice of incorporating as many processes as possible into mechanistic models. Our current

94 incomplete knowledge about ecological systems or unrepresented processes under novel

95 conditions is partly reflected in model parameters which are associated with large uncertainties.

96 Good forecasting therefore requires effective communication between process-based models and

97 data to estimate realistic model parameters and capture context-dependent ecological

98 phenomena.

99   Data-model fusion, or data-model integration, is an important step to combine models

100 with data. But previous data-model integration activities have mostly been done in an *ad hoc*

101 manner instead of being interactive. For example, data from a network of eddy covariance flux

102 tower sites across United States and Canada was compared with gross primary productivity

103 (GPP) estimated from different models (Schaefer et al., 2012). Luo and Reynolds (1999) used a

104 model to examine ecosystem responses to gradual as in the real world vs. step increases in CO_2

105 concentration as in elevated CO_2 experiments. Parton et al. (2007) parameterized CO_2 impacts in

106   an ecosystem model with data from a CO_2 experiment in Colorado. Such model-experiment

107   interactions encounter a few issues: 1) Models are not always calibrated for individual sites and,

108   therefore, not accurate; 2) It is not very effective because it is usually one-time practice without

109   many iterative processes between experimenters and modellers (Dietze et al., 2013;Lebauer et

110   al., 2013); 3) It is usually unidirectional as data is normally used to train models while the

111   guidance of model for efficient data collection is limited; and 4) It is not streamlined and could

112   not be disseminated with common practices among the research community (Lebauer et al.,

113   2013;Dietze et al., 2013;Walker et al., 2014).

114   A few research groups have developed data assimilation systems to facilitate data-model

115   integration in a systematic way. For example, data-model integration systems, such as the Data

116   Assimilation Research Testbed - DART (Anderson et al., 2009) and the Carbon Cycle Data

117   Assimilation Systems - CCDAS (Scholze et al., 2007;Peylin et al., 2016), combine various data

118   streams (e.g., FLUXNET data, satellite data and inventory data) with process-based models

119   through data assimilation algorithms such as the Kalman filter (Anderson et al., 2009) and

120   variational methods (Peylin et al., 2016). These data assimilation systems automate model

121   parameterization and provide an avenue to systematically improve models through combining as

122   much data as possible. Data-informed model improvements normally happen after the ending of

123   a field experiment and the interactive data-model integration is limited as feedbacks from models

124   to ongoing experimental studies are not adequately realised. In addition, wide applications of

125   these data assimilation systems in ecological forecasting are constrained by limited user

126   interactions with its steep learning curve to understand these systems, especially for

127   experimenters who have limited training in modelling.

128   The web-based technology facilitates interactions. Web-based modelling, which provides

129   user-friendly interfaces to run models in the background, is usually supported by the scientific

130   workflow, the sequence of processes through which a piece of work passes from initiation to

131   completion. For example, TreeWatch.Net has recently been developed to make use of high

132   precision individual tree monitoring data to parameterize process based tree models in real-time

133   and to assess instant tree hydraulics and carbon status with online result visualization (Steppe et

134   al., 2016). Although the web portal of TreeWatch.Net is currently limited to the purpose of

135   visualization, it broadens the application of data-model integration and strengthens the

136   interaction between modelling researches and the general public. The Predictive Ecosystem

137   Analyzer (PEcAn) is a scientific workflow that wraps around different ecosystem models and

138   manages the flows of information coming in and out of the model (Lebauer et al., 2013). PEcAn

139   enables web-based model simulations. Such a workflow has advantages, for example, making

140   ecological modelling and analysis convenient, transparent, reproducible and adaptable to new

141   questions (Lebauer et al., 2013), and encouraging user-model interactions. PEcAn uses the

142   Bayesian meta-analysis to synthesize plant trait data to estimate model parameters and associated

143   uncertainties, i.e., the prior information for process-based models. Parameter uncertainties are

144   propagated to model uncertainties and displayed as outputs. It is still not fully interactive in the

145   way that states are not updated iteratively according to observations and the web-based data

146   assimilation and then ecological forecasting have not yet been fully realised.

147   The iterative model-data integration provides an approach to constantly improve

148   ecological forecasting and is an important step especially in realising the near real-time

149   ecological forecasting. Instead of projecting into future through assimilating observations only

150   once, the iterative forecasting constantly updates forecasting along with ongoing new data

151 streams or/and improved models. Forecasting is likely to be improved unidirectionally in which

152 either only models are updated through observations, or only data collections/field

153 experimentations are improved according to theoretical/model information, but not both.

154 Ecological forecasting can also be bidirectionally improved so that both models and field

155 experimentations are optimized hand in hand over time. Although the bidirectional case is rare in

156 ecological forecasting, the unidirectional iterative forecasting has been reported. One excellent

157 example of forecasting through dynamically and repeatedly integrating data with models is from

158 infectious disease studies (Niu et al., 2014;Ong et al., 2010). Dynamics of infectious diseases are

159 traditionally captured by Susceptible-Infected-Removed (SIR) models. In the forecasting of the

160 Singapore H1N1-2009 infections, SIR model parameters and the number of individuals in each

161 state were updated daily, combining data renewed from local clinical reports. The evolving of the

162 epidemic related parameters and states were captured through iteratively assimilating

163 observations to inform forecasting. As a result, the model correctly forecasted the timing of the

164 peak and declining of the infection ahead of time. Iterative forecasting dynamically integrates

165 data with model and makes best use of both data and theoretical understandings of ecological

166 processes.

167   The aim of this paper is to present a fully interactive platform, a web-based Ecological

168 Platform for Assimilating Data into models (EcoPAD (v1.0)), to best inform ecological

169 forecasting. The interactive feature of EcoPAD (v1.0) is reflected in the iterative model updating

170 and forecasting through dynamically integrating models with new observations, bidirectional

171 feedbacks between experimenters and modellers, and flexible user-model communication

172 through web-based simulation, data assimilation and forecasting. Such an interactive platform

173 provides the infrastructure to effectively integrate available resources, from both models and

174   data, modellers and experimenters, scientists and the general public, to improve scientific

175   understanding of ecological processes, to boost ecological forecasting practice and transform

176   ecology towards quantitative forecasting.

177   In the following sections, we first describe the system design and major components of

178   EcoPAD (v1.0). We then use the Spruce and Peatland Responses Under Climatic and

179   Environmental change (SPRUCE) experiment (Hanson et al., 2017) as a testbed to elaborate the

180   functionality and new opportunities brought by the platform. We finally discuss implications of

181   EcoPAD (v1.0) for better ecological forecasting.

182

183   **2 EcoPAD (v1.0): system design and components**

184   **2.1 General description: web-based data assimilation and forecast**

185   EcoPAD (v1.0) (https://ecolab.nau.edu/ecopad_portal/) focuses on linking ecological

186   experiments/data with models and allows easily accessible and reproducible data-model

187   integration with interactive web-based simulation, data assimilation and forecast capabilities.

188   Specially, EcoPAD (v1.0) enables the automated near time ecological forecasting which works

189   hand-in-hand between modellers and experimenters and updates periodically in a manner similar

190   to the weather forecasting. The system is designed to streamline web request-response, data

191   management, modelling, prediction and visualization to boost the overall throughput of

192   observational data, promote data-model communication, inform ecological forecasting and

193   improve scientific understanding of ecological processes (see Supplement for detailed

194   functionalities of EcoPAD (v1.0)).

195   To realise such data-informed ecological forecasting, the essential components of

196   EcoPAD (v1.0) include experiments/data, process-based models, data assimilation techniques

9

197   and the scientific workflow (Figures 1-3). The scientific workflow of EcoPAD (v1.0) that wraps

198   around ecological models and data assimilation algorithms acts to move datasets in and out of

199   structured and catalogued data collections (metadata catalog) while leaving the logic of the

200   ecological models and data assimilation algorithms untouched (Figures 1, 3). Once a user makes

201   a request through the web browser or command line utilities, the scientific workflow takes

202   charge of triggering and executing corresponding tasks, be it pulling data from a remote server,

203   running a particular ecological model, automating forecasting or making the result easily

204   understandable to users (Figures 1, 3). With the workflow, the system is agnostic to operation

205   system, environment and programming language and is built to horizontally scale to meet the

206   demands of the model and the end user community.

207

208   **2.2 Components**

209   **2.2.1 Data**

210   Data is an important component of EcoPAD (v1.0) and EcoPAD (v1.0) offers systematic data

211   management to digest diverse data streams. The 'big data' ecology generates a large volume of

212   very different datasets across various scales (Mouquet et al., 2015;Hampton et al., 2013). These

213   datasets might have high temporal resolutions, such as those from real time ecological sensors, or

214   the display of spatial information from remote sensing sources and data stored in the geographic

215   information system (GIS). These datasets may also include, but are not limited to, inventory data,

216   laboratory measurements, FLUXNET databases or from long-term ecological networks

217   (Baldocchi et al., 2001;Johnson et al., 2010;Robertson et al., 2012) . Such data contain

218   information related to environmental forcing (e.g., precipitation, temperature and radiative

219   forcing), site characteristics (e.g., soil texture and species composition) and biogeochemical

220   information. Datasets in EcoPAD (v1.0) are derived from other research projects in comma

221   separated value files or other loosely structured data formats. These datasets are first described

222   and stored with appropriate metadata via either manual operation or scheduled automation from

223   sensors. Each project has a separate folder where data are stored. Data are generally separated

224   into two categories. One is used as boundary conditions for modelling and the other category is

225   related to observations that are used for data assimilation. Scheduled sensor data are appended to

226   existing data files with prescribed frequency. Attention is then spent on how the particular

227   dataset varies over space (x, y) and time (t). When the spatiotemporal variability is understood, it

228   is then placed in metadata records that allow for query through its scientific workflow.

229   **2.2.2 Ecological models**

230   Process-based ecological model is another essential component of EcoPAD (Figure 1). In

231   this paper, the Terrestrial ECOsystem (TECO) model is applied as a general ecological model for

232   demonstration purposes since the workflow and data assimilation system of EcoPAD (v1.0) are

233   relatively independent on the specific ecological model. Linkages among the workflow, data

234   assimilation system and ecological model are based on messaging. For example, the data

235   assimilation system generates parameters that are passed to ecological models. The state

236   variables simulated from ecological models are passed back to the data assimilation system.

237   Models may have different formulations. As long as they take in the same parameters and

238   generate the same state variables, they are functionally identical from the "eye" of the data

239   assimilation system.

240   TECO simulates ecosystem carbon, nitrogen, water and energy dynamics (Weng and

241   Luo, 2008;Shi et al., 2016). The original TECO model has 4 major submodules (canopy, soil

242   water, vegetation dynamics and soil carbon/nitrogen) and is further extended to incorporate

243   methane biogeochemistry and snow dynamics (Huang et al., 2017;Ma et al., 2017). As in the

244   global land surface model CABLE (Wang et al., 2010;Wang and Leuning, 1998), canopy

245   photosynthesis that couples surface energy, water and carbon fluxes is based on a two-big-leaf

246   model (Wang and Leuning, 1998). Leaf photosynthesis and stomatal conductance are based on

247   the common scheme from Farquhar et al. (1980) and Ball et al. (1987) respectively.

248   Transpiration and associated latent heat losses are controlled by stomatal conductance, soil water

249   content and the rooting profile. Evaporation losses of water are balanced between the soil water

250   supply and the atmospheric demand which is based on the difference between saturation vapor

251   pressure and the actual atmospheric vapor pressure. Soil moisture in different soil layers is

252   regulated by water influxes (e.g., precipitation and percolation) and effluxes (e.g., transpiration

253   and runoff). Vegetation dynamic tracks processes such as growth, allocation and phenology. Soil

254   carbon/nitrogen module tracks carbon and nitrogen through processes such as litterfall, soil

255   organic matter (SOM) decomposition and mineralization. SOM decomposition modelling

256   follows the general form of the Century model (Parton et al., 1988) as in most Earth system

257   models. SOM is divided into pools with different turnover times (the inverse of decomposition

258   rates) which are modified by environmental factors such as the soil temperature and moisture.

259   **2.2.3 Data assimilation**

260   Data assimilation is growing in importance as the process based ecological models,

261   despite largely simplifying the real systems, are in great need to be complex enough to address

262   sophisticate ecological issues. These ecological issues are composed of an enormous number of

263   biotic and abiotic factors interacting with each other. Data assimilation techniques provide a

264   framework to combine models with data to estimate model parameters (Shi et al., 2016), test

265   alternative ecological hypotheses through different model structures (Liang et al., 2015), assess

266   information content of datasets (Weng and Luo, 2011), quantify uncertainties (Zhou et al.,

267   2012;Weng et al., 2011;Keenan et al., 2012), derive emergent ecological relationships (Bloom et

268   al., 2016), identify model errors and improve ecological predictions (Luo et al., 2011b) (Figure

269   2). Under the Bayesian paradigm, data assimilation techniques treat the model structure, initial

270   and parameter values as priors that represent our current understanding of the system. As new

271   information from observations or data becomes available, model parameters and state variables

272   can be updated accordingly. The posterior distributions of estimated parameters or state variables

273   are imprinted with information from both the model and the observation/data as the chosen

274   parameters act to reduce mismatches between observations and model simulations. Future

275   predictions benefit from such constrained posterior distributions through forward modelling

276   (Figure S1). As a result, the probability density function of predicted future states through data

277   assimilation normally has a narrower spread than that without data assimilation when everything

278   else is equal (Niu et al., 2014;Luo et al., 2011b;Weng and Luo, 2011).

279   EcoPAD (v1.0) is open to different data assimilation techniques since the scientific

280   workflow of EcoPAD (v1.0) is independent on the specific data assimilation algorithm. For

281   demonstration, the Markov chain Monte Carlo (MCMC) (Xu et al., 2006) is described in this

282   study.

283   MCMC is a class of sampling algorithms to draw samples from a probability distribution

284   obtained through constructed Markov Chain to approximate the equilibrium distribution. The

285   Bayesian based MCMC method takes into account various uncertainty sources which are crucial

286   in interpreting and delivering forecasting results (Clark et al., 2001). In the application of

287   MCMC, the posterior distribution of a parameter for given observations is proportional to the

288   prior distribution of that parameter and the likelihood function which is linked to the fit/match

289   (or cost function) between model simulations and observations. EcoPAD (v1.0) currently adopts

290   a batch mode, that is, the cost function is treated as a single function to be minimized and

291   different observations are standardized by their corresponding standard deviations (Xu et al.,

292   2006). For simplicity, we assume uniform distributions in priors, and Gaussian or multivariate

293   Gaussian distributions in observational errors, which can be operationally expanded to other

294   specific distribution forms depending on the available information. Detailed description is

295   available in Xu et al. (2006).

296   **2.2.4 Scientific workflow**

297   EcoPAD (v1.0) relies on its scientific workflow to interface with ecological models and

298   data assimilation algorithms, manage diverse data streams, automates iterative ecological

299   forecasting in response to various user requests. Workflow is a relatively new concept in the

300   ecology literature but essential to realise real or near-real time forecasting. Thus, we describe it

301   in detail below. The essential components of the scientific workflow of EcoPAD (v1.0) include

302   the metadata catalog, web application-programming interface (API), the asynchronous task/job

303   queue (Celery) and the container-based virtualization platform (Docker). The workflow system

304   of EcoPAD (v1.0) also provides structured result access and visualization.

305   **2.2.4.1 Metadata catalog and data management**

306   Datasets can be placed and queried in EcoPAD (v1.0) via a common metadata catalog

307   which allows for effective management of diverse data streams. Calls for good management of

308   current large and heterogeneous ecological datasets are common (Vitolo et al., 2015;Michener

309   and Jones, 2012;Ellison, 2010). Kepler (Ludascher et al., 2006) and the Analytic Web (Osterweil

310   et al., 2010) are two example systems that endeavour to provide efficient data management

311   through the storage of metadata including clear documentation of data provenance. Similarly to

312 these systems, EcoPAD (v1.0) takes advantage of modern information technology, especially the

313 metadata catalog, to manage diverse data streams. The EcoPAD (v1.0) metadata schema includes

314 description of the data product, security, access pattern, and timestamp of last metadata update

315 *etc*. We use MongoDB (https://www.mongodb.com/), a NoSQL database technology, to manage

316 heterogeneous datasets to make the documentation, query and storage fast and convenient.

317 Through MongoDB, measured datasets can be easily fed into ecological models for various

318 purposes such as to initialize the model, calibrate model parameters, evaluate model structure

319 and drive model forecast. For datasets from real time ecological sensors that are constantly

320 updating, EcoPAD (v1.0) is set to automatically fetch new data streams with adjustable

321 frequency according to research needs.

322 **2.2.4.2 Web API, asynchronous task queue and docker**

323 The RESTful application-programming interface (API) which can deliver data to a wide

324 variety of applications is the gateway of EcoPAD (v1.0) and enables a wide array of user-

325 interfaces and data-dissemination activities. Once a user makes a request, such as through

326 clicking on relevant buttons from a web browser, the request is passed through the

327 Representational State Transfer (i.e., RESTful) API to trigger specific tasks. The RESTful API

328 bridges the talk between the client (e.g., a web browser or command line terminal) and the server

329 (Figure 3). The API exploits the full functionality and flexibility of the HyperText Transfer

330 Protocol (HTTP), such that data can be retrieved and ingested from the EcoPAD (v1.0) through

331 the use of simple HTTP headers and verbs (e.g., GET, PUT, POST, *etc*.). Hence, a user can

332 incorporate summary data from EcoPAD (v1.0) into a website with a single line of html code.

333 Users will also be able to access data directly through programming environments like R, Python

334 and Matlab. Simplicity, ease of use and interoperability are among the main advantages of this

335 API which enables web-based modelling.

336   Celery (https://github.com/celery/celery) is an asynchronous task/job queue that runs in

337 the background (Figure 3). The task queue (i.e., Celery) is a mechanism used to distribute work

338 across work units such as threads or machines. Celery communicates through messages, and

339 EcoPAD (v1.0) takes advantage of the RabbitMQ (https://www.rabbitmq.com/) to manage

340 messaging. After the user submits a command, the request or message is passed to Celery via the

341 RESTful API. These messages may trigger different tasks, which include, but not limited to, pull

342 data from a remote server where original measurements are located, access data through

343 metadata catalog, run model simulation with user specified parameters, conduct data assimilation

344 which recursively updates model parameters, forecast future ecosystem status and post-process

345 of model results for visualization. The broker inside Celery receives task messages and handles

346 out tasks to available Celery workers which perform the actual tasks (Figure 3). Celery workers

347 are in charge of receiving messages from the broker, executing tasks and returning task results.

348 The worker can be a local or remote computation resource (e.g., the cloud) that has connectivity

349 to the metadata catalog. Workers can be distributed into different information technology (IT)

350 infrastructures, which makes EcoPAD (v1.0) workflow expandable. Each worker can perform

351 different tasks depending on tools installed in each worker. And one task can also be distributed

352 into different workers. In such a way, EcoPAD (v1.0) workflow enables parallelization and

353 distributed computation of actual modelling tasks across various IT infrastructures, and is

354 flexible in implementing additional computational resources by connecting additional workers.

355   Another key feature that makes EcoPAD (v1.0) easily portable and scalable among

356 different operation systems is the utilization of the container-based virtualization platform, the

357   docker (https://www.docker.com/). Docker can run many applications which rely on different

358   libraries and environments on a single kernel with its lightweight containerization. Tasks that

359   execute TECO in different ways are wrapped inside different docker containers that can "talk"

360   with each other. Each docker container embeds the ecosystem model into a complete filesystem

361   that contains everything needed to run an ecosystem model: the source code, model input, run

362   time, system tools and libraries. Docker containers are both hardware-agnostic and platform-

363   agnostic, and they are not confined to a particular language, framework or packaging system.

364   Docker containers can be run from a laptop, workstation, virtual machine, or any cloud compute

365   instance. This is done to support the widely varied number of ecological models running in

366   various languages (e.g., Matlab, Python, Fortran, C and C++) and environments. In addition to

367   wrap the ecosystem model into a docker container, software applied in the workflow, such as the

368   Celery, Rabbitmq and MongoDB, are all lightweight and portable encapsulations through docker

369   containers. Therefore, the entire EcoPAD (v1.0) is readily portable and applicable in different

370   environments.

371   **2.2.4.3 Structured result access and visualization**

372   EcoPAD (v1.0) enables structured result storage, access and visualization to track and

373   analyse data-model fusion practice. Upon the completion of the model task, the model wrapper

374   code calls a post processing call-back function. This call-back function allows for model specific

375   data requirements to be added to the model result repository. Each task is associated with a

376   unique task ID and model results are stored within the local repository that can be queried by the

377   unique task ID. The store and query of model results are realised via the MongoDB and RESTful

378   API (Figure 3). Researchers are authorized to review and download model results and parameters

379   submitted for each model run through a web accessible URL (link). EcoPAD (v1.0) webpage

380   also displays a list of historical tasks (with URL) performed by each user. All current and

381   historical model inputs and outputs are available to download, including the aggregated results

382   produced for the graphical web applications. In addition, EcoPAD (v1.0) also provides a task

383   report that contains all-inclusive recap of submitted parameters, task status, and model outputs

384   with links to all data and graphical results for each task. Such structured result storage and access

385   make sharing, tracking and referring to modelling studies instant and clear.

386

387   **3 EcoPAD (v1.0) performance at testbed - SPRUCE**

388   **3.1 SPRUCE project overview**

389   EcoPAD (v1.0) is being applied to the Spruce and Peatland Responses Under Climatic

390   and Environmental change (SPRUCE) experiment located at the USDA Forest Service Marcell

391   Experimental Forest (MEF, 47°30.476' N, 93°27.162' W) in northern Minnesota (Kolka et al.,

392   2011). SPRUCE is an ongoing project that focuses on long-term responses of northern peatland

393   to climate warming and increased atmospheric CO_2 concentration (Hanson et al., 2017). At

394   SPRUCE, ecologists measure various aspects of responses of organisms (from microbes to trees)

395   and ecological functions (carbon, nutrient and water cycles) to a warming climate. One of the

396   key features of the SPRUCE experiments is the manipulative deep soil/peat heating (0-3 m) and

397   whole ecosystem warming treatments (peat + air warmings) which include tall trees (> 4 m)

398   (Hanson et al., 2017). Together with elevated atmospheric CO_2 treatments, SPRUCE provides a

399   platform for exploring mechanisms controlling the vulnerability of organisms, biogeochemical

400   processes and ecosystems in response to future novel climatic conditions. The SPRUCE peatland

401   is especially sensitive to future climate change and also plays an important role in feeding back

402   to future climate change through greenhouse gas emissions as it stores a large amount of soil

403   organic carbon. Vegetation in the SPRUCE site is dominated by *Picea mariana* (black spruce)

404   and *Sphagnum spp* (peat moss). The studied peatland also has an understory which include

405   ericaceous and woody shrubs. There are also a limited number of herbaceous species. The whole

406   ecosystem warming treatments include a large range of both aboveground and belowground

407   temperature manipulations (ambient, control plots of + 0 °C, +2.25 °C, +4.5 °C, +6.75 °C and +9

408   °C) in large 115 m^2 open-topped enclosures with elevated CO_2 manipulations (+0 or +500 ppm).

409   The difference between ambient and +0 °C treatment plots is the open-topped and controlled-

410   environment enclosure.

411   The SPRUCE project generates a large variety of observational datasets that reflect

412   ecosystem dynamics from different scales and are available from the project webpage

413   (https://mnspruce.ornl.gov/) and FTP site (ftp://sprucedata.ornl.gov/). These datasets come from

414   multiple sources: half hourly automated sensor records, species surveys, laboratory

415   measurements, laser scanning images *etc*. Involvements of both modelling and experimental

416   studies in the SPRUCE project create the opportunity for data-model communication. Datasets

417   are pulled from SPRUCE archives and stored in the EcoPAD (v1.0) metadata catalog for running

418   the TECO model, conducting data-model fusion or forecasting. The TECO model has been

419   applied to simulate and forecast carbon dynamics with productions of CO_2 and CH_4 from

420   different carbon pools, soil temperature response, snow depth and freeze-thaw cycles at the

421   SRPUCE site (Jiang et al., 2018;Huang et al., 2017;Ma et al., 2017).

422

423   **3.2 EcoPAD-SPRUCE web portal**

424   We assimilate multiple streams of data from the SPRUCE experiment to the TECO

425   model using the MCMC algorithm, and forecast ecosystem dynamics in both near time and for

426   the next 10 years. Our forecasting system for SPRUCE is available at

427   https://ecolab.nau.edu/ecopad_portal/. From the web portal, users can check our current near-

428   and long-term forecasting results, conduct model simulation, data assimilation and forecasting

429   runs, and analyse/visualize model results. Detailed information about the interactive web portal

430   is provided in the supplementary information.

431   **3.3 Near time ecosystem forecasting and feedback to experimenters**

432   As part of the forecasting functionality, EcoPAD-SPRUCE automates the near time

433   (weekly) forecasting with continuously updated observations from SPRUCE experiments (Figure

434   4). We set up the system to automatically pull new data streams every Sunday from the SPRUCE

435   FTP site that holds observational data and update the forecasting results based on new data

436   streams. Updated forecasting results for the next week are customized for the SPRUCE

437   experiments with different manipulative treatments and displayed in the EcoPAD-SPRUCE

438   portal. At the same time, these results are sent back to SPRUCE communities and displayed

439   together with near-term observations for experimenter's reference.

440   **3.4 New approaches to ecological studies towards better forecasting**

441   **3.4.1 Case 1: Interactive communications among modellers and experimenters**

442   EcoPAD-SPRUCE provides a platform to stimulate interactive communications between

443   modellers and experimenters through the loop of prediction-question-discussion-adjustment-

444   prediction (Figure 4). We illustrate how the prediction-question-discussion-adjustment-

445   prediction cycle and stimulation of modeller-experimenter communication improves ecological

446   predictions through one episode during the study of the relative contribution of different

447   pathways to methane emissions. An initial methane model was built upon information (e.g., site

448   characteristics and environmental conditions) provided by SPRUCE field scientists, taking into

20

449   account important processes in methane dynamics, such as production, oxidation and emissions

450   through three pathways (i.e., diffusion, ebullition and plant-mediated transportation). The model

451   was used to predict relative contributions of different pathways to overall methane emissions

452   under different warming treatments after being constrained by measured surface methane fluxes.

453   Initial forecasting results which indicated a strong contribution from ebullition under high

454   warming treatments were sent back to the SPRUCE group. Experimenters doubted about such a

455   high contribution from the ebullition pathway and a discussion was stimulated. It is difficult to

456   accurately distinguish the three pathways from field measurements. Field experimenters

457   provided potential avenues to extract measurement information related to these pathways, while

458   modellers examined model structure and parameters that may not be well constrained by

459   available field information. Detailed discussion is provided in Table 1. After extensive

460   discussion, several adjustments were adopted as a first step to move forward. For example, the

461   three-porosity model that was used to simulate the diffusion process was replaced by the

462   Millington-Quirk model to more realistically represent methane diffusions in peat soil; the

463   measured static chamber methane fluxes were also questioned and scrutinized more carefully to

464   clarify that they did not capture the episodic ebullition events. Measurements such as these

465   related to pore water gas data may provide additional inference related to ebullition. The updated

466   forecasting is more reasonable than the initial results although more studies are in need to

467   ultimately quantify methane fluxes from different pathways.

468   **3.4.2 Case 2: Acclimation of ecosystem carbon cycling to experimental manipulations**

469   As a first step, CH_4 static chamber flux measurements were assimilated into TECO to

470   assess potential acclimation phenomena during methane production under 5 warming treatments

471   (+0, +2.25, +4.5, +6.75, +9 °C). Initial results indicated a reduction in both the $CH_4{:}CO_2$ ratio

472 and the temperature sensitivity of methane production based on their posterior distributions

473 (Figure 5). The mean $CH_4:CO_2$ ratio decreased from 0.675 (+0 $^\circ$C treatment) to 0.505 (+9 $^\circ$C),

474 while the temperature sensitivity (Q_{10}) for CH_4 production decreased from 3.33 (+0 $^\circ$C) to 1.22

475 (+9 $^\circ$C treatment). Such shifts quantify potential acclimation of methane production to warming

476 and future climate warming is likely to have a smaller impact on emission than most of current

477 predictions that do not take into account of acclimation.

478   Despite these results are preliminary as more relevant datasets are under collection with

479 current ongoing warming manipulations and measurements, assimilating observations through

480 EcoPAD (v1.0) provides a quantitative approach to timely assess acclimation through time.

481 Melillo et al. (2017) revealed that the thermal acclimation of the soil respiration in the Harvard

482 Forest is likely to be phase (time) dependent during their 26-year soil warming experiment.

483 EcoPAD (v1.0) provides the possibility in tracing the temporal path of acclimation with its

484 streamlined structure and archive capacity. Shi et al. (2015) assimilated carbon related

485 measurements in a tallgrass prairie into the TECO model to study acclimation after 9-years

486 warming treatments. They revealed a reduction in the allocation of GPP to shoot, the turnover

487 rates of the shoot and root carbon pools, and an increase in litter and fast carbon turnovers in

488 response to warming treatments. Similarly, as time goes on, the SPRUCE experiment will

489 generate more carbon cycling related datasets under different warming and CO_2 treatments,

490 which can be mounted to EcoPAD (v1.0) to systematically quantify acclimations in carbon

491 cycling through time in the future.

492 **3.4.3 Case 3: Partitioning of uncertainty sources**

493   Uncertainties in ecological studies can come from observations (include forcing that

494 drives the model), different model structures to represent the real world and the specified model

495 parameters (Luo et al., 2016). Previous studies tended to focus on one aspect of the uncertainty

496 sources instead of disentangling the contribution from different sources. For example, the model

497 intercomparison projects (MIPs), such as TRENDY, focus on uncertainty caused by different

498 model structures with prescribed external forcing (Sitch et al., 2008). Keenan et al. (2012) used

499 data assimilation to constrain parameter uncertainties in projecting Harvard forest carbon

500 dynamics. Ahlstrom et al. (2012) forced one particular vegetation model by 18 sets of forcings

501 from climate models of the Coupled Model Intercomparison Project Phase 5 (CMIP5), while the

502 parameter or model structure uncertainty is not taken into account.

503   EcoPAD (v1.0) is designed to provide a thorough picture of uncertainties from multiple

504 sources especially in carbon cycling studies. Through focusing on multiple instead of one source

505 of uncertainty, ecologists can allocate resources to areas that cause relative high uncertainty.

506 Attribution of uncertainties in EcoPAD (v1.0) will rely on an ensemble of ecosystem models, the

507 data assimilation system and climate forcing with quantified uncertainty. *Jiang et al.* [20Jiang et

508 al. (2018) focused specifically on the relative contribution of parameter uncertainty vs. climate

509 forcing uncertainty in forecasting carbon dynamics at the SPRUCE site. Through assimilating

510 the pre-treatment measurements (2011-2014) from the SPRUCE experiment, Jiang et al. (2018)

511 estimated uncertainties of key parameters that regulate the peatland carbon dynamics. Combined

512 with the stochastically generated climate forcing (e.g., precipitation and temperature), Jiang et al.

513 (2018) found external forcing resulted in higher uncertainty than parameters in forecasting

514 carbon fluxes, but caused lower uncertainty than parameters in forecasting carbon pools.

515 Therefore, more efforts are required to improve forcing measurements for studies that focus on

516 carbon fluxes (e.g., GPP), while reductions in parameter uncertainties are more important for

517 studies in carbon pool dynamics. Despite Jiang et al. (2018) does not quantify model structure

518 uncertainty, the project of incorporating multiple models inside EcoPAD (v1.0) is in progress,

519 and future uncertainty assessment will benefit from EcoPAD (v1.0) with its systematically

520 archived model simulation, data assimilation and forecasting.

521 **3.4.4 Case 4: Improving biophysical estimation for better ecological prediction**

522 Carbon cycling studies can also benefit from EcoPAD (v1.0) through improvements in

523 biophysical estimation. Soil environmental condition is an important regulator of belowground

524 biological activities and also feeds back to aboveground vegetation growth. Biophysical

525 variables such as soil temperature, soil moisture, ice content and snow depth, are key predictors

526 of ecosystem dynamics. After constraining the biophysical module by detailed monitoring data

527 from the SPRUCE experiment through the data assimilation component of EcoPAD (v1.0),

528 Huang et al. (2017) forecasted the soil thermal dynamics under future conditions and studied the

529 responses of soil temperature to hypothetical air warming. This study emphasized the importance

530 of accurate climate forcing in providing robust thermal forecast. In addition, Huang et al. (2017)

531 revealed non-uniform responses of soil temperature to air warming. Soil temperature responded

532 stronger to air warming during summer compared to winter. And soil temperature increased

533 more in shallow soil layers compared to deep soils in summer in response to air warming.

534 Therefore, extrapolating of manipulative experiments based on air warming alone may not

535 reflect the real temperature sensitivity of SOM if soil temperature is not monitored. As robust

536 quantification of environmental conditions is known to be a first step towards better

537 understanding of ecological process, improvement in soil thermal predictions through EcoPAD

538 (v1.0) data assimilation system is helpful in telling apart biogeochemical responses from

539 environmental uncertainties and also in providing field ecologists beforehand key environmental

540 conditions.

3.4.5 Case 5: How do updated model and data contribute to reliable forecasting?

Through constantly adjusted model and external forcing according to observations and weekly archived model parameter, model structure, external forcing and forecasting results, the contribution of model and data updates can therefore be tracked through comparing forecasted vs. realised simulations. For example, Figure 6 illustrates how updated external forcing (compared to stochastically generated forcing) and shifts in ecosystem state variables shape ecological predictions. "updated" means the real meteorological forcing monitored from the weather station. We use stochastically generated forcing to represent future meteorological conditions. Future precipitation and air temperature were generated by vector autoregression using historical dataset (1961–2014) monitored by the weather station. PAR, relative humidity and wind speed were randomly sampled from the joint frequency distribution at a given hour each month. Detailed information on weather forcing is available in *Jiang et al.* [20Jiang et al. (2018). Similarly as in other EcoPAD-SPURCE case studies, TECO is trained through data assimilation with observations from 2011-2014 and is used to forecast GPP and total soil organic carbon content at the beginning of 2015. For demonstrating purpose, Figure 6 only shows 3 series of forecasting results instead of updates from every week. Series 1 (S1) records forecasted GPP and soil carbon with stochastically generated weather forcing from January 2015-December 2024 (Figure 6a,b cyan). Series 2 (S2) records simulated GPP and soil carbon with observed (updated) climate forcing from January 2015 to July 2016 and forecasted GPP and soil carbon with stochastically generated forcing from August 2016 - December 2024 (Figure 6a,b red). Similarly, the stochastically generated forcing in Series 3 (S3) starts from January 2017 (Figure 6a,b blue). For each series, predictions were conducted with randomly sampled parameters from the posterior distributions and stochastically generated forcing. We displayed 100 mean values (across an

564   ensemble of forecasts with different parameters) corresponding to 100 forecasts with

565   stochastically generated forcing.

566   GPP is highly sensitive to climate forcing. The differences between the updated (S2, 3)

567   and initial forecasts (S1) reach almost 800 gC m^{-2} year^{-1} (Figure 6c). The discrepancy is strongly

568   dampened in the following 1-2 years. The impact of updated forecasts is close to 0 after

569   approximately 5 years. However, soil carbon pool shows a different pattern. Soil carbon pool is

570   increased by less than 150 gC m^{-2}, which is relative small compared to the carbon pool size of *ca.*

571   62000 gC m^{-2}. The impact of updated forecasts grows with time and reaches the highest at the

572   end of the simulation year 2024. GPP is sensitive to the immediate change in climate forcing

573   while the updated ecosystem status (or initial value) has minimum impact in the long-term

574   forecast of GPP. The impact of updated climate forcing is relatively small for soil carbon

575   forecasts during our study period. Soil carbon is less sensitive to the immediate change of

576   climate compared to GPP. However, the alteration of system status affects soil carbon forecast

577   especially in a longer time scale.

578   Since we are archiving updated forecasts every week, we can track the relative

579   contribution of ecosystem status, forcing uncertainty and parameter distributions to the overall

580   forecasting patterns of different ecological variables and how these patterns evolve in time. In

581   addition, as growing observations of ecological variables (e.g., carbon fluxes and pool sizes)

582   become available, it is feasible to diagnose key factors that promote robust ecological forecasting

583   through comparing the archived forecasts vs. observation and analysing archives of model

584   parameters, initial values and climate forcing *etc*.

585

586   **4 Discussion**

4.1 The necessity of interactive infrastructure to realise ecological forecasting

Interactions enable exchanging and extending of information so as to benefit from collective knowledge. For example, manipulative studies will have a much broader impact if the implications of their results can be extended from the regression between environmental variable and ecosystem response, such as be integrated into an ecosystem model through model-data communication. Such an approach will allow gaining information about the processes responsible for ecosystem's response, constraining models, and making more reliable predictions. Going beyond common practice of model-data assimilation from which model updating lags far behind observations, EcoPAD (v1.0) enables iterative model updating and forecasting through dynamically integrating models with new observations in near real-time. This near real-time interactive capacity relies on its scientific workflow that automates data management, model simulation, data simulation and result visualization. The system design encourages thorough interactions between experimenters and modellers. Forecasting results from SPRUCE were timely shared among research groups with different background through the web interface. Expertise from different research groups was integrated to improve a second round of forecasting. Again, thanks to the workflow, new information or adjustment is incorporated into forecasting efficiently, making the forecasting system fully interactive.

We also benefit from the interactive EcoPAD (v1.0) platform to broaden user-model interactions and to broadcast forecasting results. Learning about the ecosystem models and data-model fusion techniques may lag one's productivity and even discourage learning the modelling techniques because of their complexity and long learning curve. Because EcoPAD (v1.0) can be accessed from a web browser and does not require any coding from the user's side, the time lag between learning the model structure and obtaining model-based results for one's study is

610   minimal, which opens the door for non-modeller groups to "talk" with models. The online

611   storage of one's results lowers the risk of data loss. The results of each model run can be easily

612   tracked and shared with its unique ID and web address. In addition, the web-based workflow also

613   saves time for experts with automated model running, data assimilation, forecasting, structured

614   result access and instantaneous graphic outputs, bringing the possibility for thorough exploration

615   of more essence part of the system. The simplicity in use of EcoPAD (v1.0) at the same time

616   may limit their access to the code and lowers the flexibility. Flexibility for users with higher

617   demands, for example, those who wanted to test alternative data assimilation methods, use a

618   different carbon cycle model, change the number of calibrated parameters, include the

619   observations for other variables, is provided through the GitHub repository

620   (https://github.com/ou-ecolab). This GitHub repository contains code and instruction for

621   installing, configuring and controlling the whole system, users can adapt the workflow to wrap

622   their own model based on his or her needs. On one hand the web-based system with open source

623   broadens the user community. On the other hand, it increases the risk of misuse and

624   misinterpretation. We encourage users to be critical and consult system developers to avoid

625   inappropriate application of the system.

626   **4.2 Implications for better ecological forecasting**

627   Specifically to reliable forecasting of carbon dynamics, our initial exploration from

628   EcoPAD-SPRUCE indicates that realistic model structure, correct parameterization and accurate

629   external environmental conditions are essential. Model structure captures important mechanisms

630   that regulate ecosystem carbon dynamics. Adjustment in model structure is critical in our

631   improvement in methane forecasting. Model parameters may vary between observation sites,

632   change with time or environmental conditions (Medlyn et al., 1999;Luo et al., 2001). A static or

633   wrong parameterization misses important mechanisms (e.g., acclimation and adaptation) that

634   regulate future carbon dynamics. Not well constrained parameters, for example, caused by lack

635   of information from observational data, contribute to high forecasting uncertainty and low

636   reliability of forecasting results. Correct parameterization is especially important for long-term

637   carbon pool predictions as parameter uncertainty resulted in high forecasting uncertainty in our

638   case study (Jiang et al., 2018). Parameter values derived under the ambient condition was not

639   applicable to the warming treatment in our methane case due to acclimation. External

640   environmental condition is another important factor in carbon predictions. External

641   environmental condition includes both the external climatic forcing that is used to drive

642   ecosystem models and also the environmental condition that is simulated by ecosystem models.

643   As we showed that air warming may not proportionally transfer to soil warming, realistic soil

644   environmental information needs to be appropriately represented to predict soil carbon dynamics

645   (Huang et al., 2017). The impact of external forcing is especially obvious in short-term carbon

646   flux predictions. Forcing uncertainty resulted in higher forecasting uncertainty in carbon flux

647   compared to that from parameter uncertainty (Jiang et al., 2018). Mismatches in forecasted vs.

648   realised forcing greatly increased simulated GPP and the discrepancy diminished in the long run.

649   Reliable external environmental condition, to some extent, reduces the complexity in diagnosing

650   modelled carbon dynamics.

651   Pool-based vs. flux-based predictions are regulated differently by external forcing and

652   initial states, which indicates that differentiated efforts are required to improve short- vs. long-

653   term predictions. External forcing, which has not been well emphasized in previous carbon

654   studies, has strong impact on short-term forecasting. The large response of GPP to forecasted vs.

655   realised forcing as well the stronger forcing-caused uncertainty in GPP predictions indicate

656   correct forcing information is a key step in short-term flux predictions. In this study, we

657   stochastically generated the climate forcing based on local climatic conditions (1961-2014),

658   which is not sufficient in capturing local short-term climate variability. As a result, updated GPP

659   went outside our ensemble forecasting. On the other hand, parameters and historical information

660   about pool status are more important in long-term pool predictions. Therefore, improvement in

661   long-term pool size predictions cannot be reached by accurate climatic information alone.

662   Instead, it requires accumulation in knowledge related to site history and processes that regulate

663   pool dynamics.

664   Furthermore, reliable forecasting needs understanding of uncertainty sources in addition

665   to the future mean states. Uncertainty and complexity are major reasons that lead to the belief in

666   "computationally irreducible" and low intrinsic predictability of ecological systems (Beckage et

667   al., 2011;Coreau et al., 2010;Schindler and Hilborn, 2015). Recent advance in computational

668   statistical methods offers a way to formally accounting for various uncertainty sources in

669   ecology (Clark et al., 2001;Cressie et al., 2009). And the Bayesian approach embedded in

670   EcoPAD (v1.0) brings the opportunity to understand and communicate forecasting uncertainty.

671   Our case study revealed that forcing uncertainty is more important in flux-based predictions

672   while parameter uncertainty is more critical in pool-based predictions. Actually, how forecasting

673   uncertainty changes with time, what are the dominate contributor of forecasting uncertainty (e.g.,

674   parameter, initial condition, model structure, observation errors, forcing *etc.*), how uncertainty

675   sources interact among different components, or to what extent unconstrained parameters affect

676   forecasting uncertainty are all valuable questions that can be explored through EcoPAD (v1.0).

677   **4.3 Applications of EcoPAD (v1.0) to manipulative experiments and observation sites**

678   Broadly speaking, data-model integration stands to increase the overall precision and

679   accuracy of model-based experimentation (Luo et al., 2011b;Niu et al., 2014). Systems for which

680   data have been collected in the field and which are well represented by ecological models

681   therefore have the capacity to receive the highest benefits from EcoPAD (v1.0) to improve

682   forecasts. In a global change context, experimental manipulations including ecosystem responses

683   to changes in precipitation regimes, carbon dioxide concentrations, temperatures, season lengths,

684   and species compositional shifts can now be assimilated into ecosystem models (Shi et al.,

685   2016;Xu et al., 2006;Gao et al., 2011;Lebauer et al., 2013). Impacts of these global change

686   factors on carbon cycling and ecosystem functioning can now be measured in a scientifically

687   transparent and verifiable manner. This leads to ecosystem modelling of systems and processes

688   that can obtain levels of confidence that lend credibility with the public to the science's forward

689   progress toward forecasting and predicting (Clark et al., 2001). These are the strengths of a

690   widely-available interface devoted to data-model integration towards better forecasting.

691   The data-model integration framework of EcoPAD (v1.0) creates a smart interactive

692   model-experiment (ModEx) system. ModEx has the capacity to form a feedback loop in which

693   field experiment guides modelling and modelling influences experimental focus (Luo et al.,

694   2011a). We demonstrated how EcoPAD (v1.0) works hand-in-hand between modellers and

695   experimenters in the life-cycle of the SPRUCE project. The EcoPAD-SPRUCE system operates

696   while experimenters are making measurements or planning for future researches. Information is

697   constantly fed back between modellers and experimenters, and simultaneous efforts from both

698   parties illustrate how communications between model and data advance and shape our

699   understanding towards better forecasts during the lifecycle of a scientific project. ModEx can be

700   extended to other experimental systems to: 1, predict what might be an ecosystem's response to

31

701   treatments once experimenter selected a site and decided the experimental plan; 2, assimilate

702   data experimenters are collecting along the experiment to constrain model predictions; 3, project

703   what an ecosystem's responses may likely be in the rest of the experiment; 4, tell experimenters

704   what are those important datasets experimenters may want to collect in order to understand the

705   system; 5, periodically updates the projections; and 6, improve the models, the data assimilation

706   system, and field experiments during the process.

707   In addition to the manipulative experiments, the data assimilation system of EcoPAD

708   (v1.0) can be used for automated model calibration for FLUXNET sites or other observation

709   networks, such as the NEON and LTER (Johnson et al., 2010;Robertson et al., 2012). The

710   application of EcoPAD (v1.0) at FLUXNET, NEON or LTER sites includes three steps in

711   general. First, build the climate forcing in the suitable formats of EcoPAD (v1.0) from the

712   database of each site; Second, collect the prior information (include observations of state

713   variables) in the data assimilation system from FLUXNET, NEON or LTER sites; Third,

714   incorporate the forcing and prior information into EcoPAD (v1.0), and then run the EcoPAD

715   (v1.0) with the dynamic data assimilation system. Furthermore, facing the proposed continental

716   scale ecology study (Schimel, 2011), EcoPAD (v1.0) once properly applied could also help

717   evaluate and optimize field deployment of environmental sensors and supporting

718   cyberinfrastructure, that will be necessary for larger, more complex environmental observing

719   systems being planned in the US and across different continents.

720   **4.4 Future developments**

721   EcoPAD (v1.0) will expand as time goes on. The system is designed to incorporate

722   multiple process-based models, diverse data assimilation techniques and various ecological state

723   variables for different ecosystems. Case studies presented in earlier sections are based primarily

724   on one model. A multiple (or ensemble) model approach is helpful in tracking uncertainty

725   sources from our process understanding. With rapid evolving ecological knowledge, emerging

726   models with different hypotheses, such as the microbial-enzyme model (Wieder et al., 2013),

727   enhance our capacity in ecological prediction but can also benefit from rapid tests against data if

728   incorporated into EcoPAD (v1.0). In addition to MCMC (Braswell et al., 2005;Xu et al., 2006), a

729   variety of data assimilation techniques have been recently applied to improve models for

730   ecological forecasting, such as the EnKF (Gao et al., 2011), Genetic Algorithm (Zhou and Luo,

731   2008) and 4-d variational assimilation (Peylin et al., 2016). Future development will incorporate

732   different optimization techniques to offer users the option to search for the best model

733   parameters by selecting and comparing the possibly best method for their specific studies. We

734   focus mostly on carbon related state variables in the SPRUCE example, and the data assimilation

735   system in EcoPAD (v1.0) needs to include more observed variables for constraining model

736   parameters. For example, the NEON sites not only provide measured ecosystem CO_2 fluxes and

737   soil carbon stocks, but also resources (e.g., GPP/Transpiration for water and GPP/intercepted

738   PAR for light) use efficiency (Johnson et al., 2010).

739   　　　Researchers interested in creating their own multiple model and/or multiple assimilation

740   scheme version of EcoPAD (v1.0) can start from the GitHub repository (https://github.com/ou-

741   ecolab) where the source code of the EcoPAD (v1.0) workflow is archived. To add a new

742   variable that is not forecasted in the EcoPAD-SPRUCE example, it requires modellers and

743   experimenters to work together to understand their process-based model, observations and how

744   messaging works in the workflow of EcoPAD (v1.0) following the example of EcoPAD-

745   SPRUCE. To add a new model or a new data assimilation scheme for variables that are

746   forecasted in EcoPAD-SPRUCE, researchers need to create additional dockers and mount them

747 to the existing workflow with the knowledge of how information are passed within the workflow

748 (see Supplement for detailed information).

749   With these improvements, one goal of the EcoPAD (v1.0) is to enable the research

750 community to understand and reduce forecasting uncertainties from different sources and

751 forecast various aspects of future biogeochemical and ecological changes as data becomes

752 available. EcoPAD (v1.0) acts as a tool to link model and data, not as a substitution for neither

753 model nor data. Ecological forecasting through EcoPAD (v1.0) relies strongly on theoretical

754 (model) and empirical (data) ecological studies. Questions such as what are major factors

755 regulating temporal variability of methane emissions cannot be directly answered by EcoPAD

756 (v1.0). How to make use of EcoPAD (v1.0) to inspire breakthroughs in both theoretical and

757 empirical ecological studies worth future exploration.

758   The power of EcoPAD (v1.0) also lies in the potential service it can bring to the society.

759 Forecasting with carefully quantified uncertainty is helpful in providing support for natural

760 resource manager and policy maker (Clark et al., 2001). It is always difficult to bring the

761 complex mathematical ecosystem models to the general public, which creates a gap between

762 current scientific advance and public awareness. The web-based interface from EcoPAD (v1.0)

763 makes modelling as easy as possible without losing the connection to the mathematics behind the

764 models. It will greatly transform environmental education and encourage citizen science (Miller-

765 Rushing et al., 2012;Kobori et al., 2016) in ecology and climate change with future outreach

766 activities to broadcast the EcoPAD (v1.0) platform.

767 **5 Conclusion**

768   The fully interactive web-based Ecological Platform for Assimilating Data (EcoPAD

769 (v1.0)) into models aims to promote data-model integration towards predictive ecology through

770   bringing the complex ecosystem model and data assimilation techniques accessible to different

771   audience. It is supported by meta-databases of biogeochemical variables, libraries of modules of

772   process models, toolbox of inversion techniques and the scalable scientific workflow. Through

773   these components, it automates data management, model simulation, data assimilation,

774   ecological forecasting, and result visualization, providing an open, convenient, transparent,

775   flexible, scalable, traceable and readily portable platform to systematically conduct data-model

776   integration towards better ecological forecasting.

777   We illustrated several of its functionalities through the Spruce and Peatland Responses

778   Under Climatic and Environmental change (SPRUCE) experiment. The iterative forecasting

779   approach from EcoPAD-SPRUCE through the prediction-question-discussion-adjustment-

780   prediction cycle and extensive communication between model and data creates a new paradigm

781   to best inform forecasting. In addition to forecasting, EcoPAD (v1.0) enables interactive web-

782   based approach to conduct model simulation, estimate model parameters or state variables,

783   quantify uncertainty of estimated parameters and projected states of ecosystems, evaluate model

784   structures, and assess sampling strategies. Altogether, EcoPAD-SPRUCE creates a smart

785   interactive model-experiment (ModEx) system from which experimenters can know what an

786   ecosystem's response might be at the beginning of their experiments, constrain models through

787   collected measurements, predict ecosystem's response in the rest of the experiments, adjust

788   measurements to better understand their system, periodically update projections and improve

789   models, the data assimilation system, and field experiments.

790   Specifically to forecasting carbon dynamics, EcoPAD-SPRUCE revealed that better

791   forecasting relies on improvements in model structure, parameterization and accurate external

792   forcing. Accurate external forcing is critical for short-term flux-based carbon predictions while

35

793   right process understanding, parameterization and historical information are essential for long-

794   term pool-based predictions. In addition, EcoPAD (v1.0) provides an avenue to disentangle

795   different sources of uncertainties in carbon cycling studies and to provide reliable forecasts with

796   accountable uncertainties.

797

798   **Code availability:**

799   EcoPAD (v1.0) portal is available at https://ecolab.nau.edu/ecopad_portal/ and code is provided

800   at the GitHub repository (https://github.com/ou-ecolab).

801   **Data availability:**

802   Relevant data for this manuscript is available at the SPRUCE project webpage

803   (https://mnspruce.ornl.gov/) and the EcoPAD (v1.0) web portal

804   (https://ecolab.nau.edu/ecopad_portal/). Additional data can be requested from the

805   corresponding author.

806   **Competing interests:**

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

808   **Author contribution**:

809   Y. Huang., M. Stacy., J. Jiang and Y.Luo designed this study. Y. Huang, J. Jiang, N. Sundi and

810   M. Stacy built the workflow. J. Jiang built the simulation and data assimilation systems. S. Ma

811   tested the system and conducted researches related to methane. P. J. Hanson and D. Ricciuto

812   were in charge of observational data collection, transfer and archive. V. Saruta and C. G. Jung

813   maintain and made further improvement of the system. Y. Huang conducted the analysis and

814   drafted the manuscript. All authors discussed the results and contributed to the manuscript.

815   **Acknowledgement**:

820   **Literature Cited**

821   Ahlstrom, A., Schurgers, G., Arneth, A., and Smith, B.: Robustness and uncertainty in
822   terrestrial ecosystem carbon response to CMIP5 climate change projections, Environmental
823   Research Letters, 7, doi:10.1088/1748-9326/7/4/044008, 2012.
824   Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., and Avellano, A.: The data
825   assimilation research testbed A Community Facility, Bulletin of the American
826   Meteorological Society, 90, 1283-1296, doi:10.1175/2009bams2618.1, 2009.
827   Baldocchi, D., Falge, E., Gu, L. H., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer,
828   C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X. H., Malhi, Y., Meyers,
829   T., Munger, W., Oechel, W., U, K. T. P., Pilegaard, K., Schmid, H. P., Valentini, R., Verma, S.,
830   Vesala, T., Wilson, K., and Wofsy, S.: FLUXNET: A new tool to study the temporal and spatial
831   variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities,
832   Bulletin of the American Meteorological Society, 82, 2415-2434, doi:10.1175/1520-
833   0477(2001)082<2415:fantts>2.3.co;2, 2001.
834   Bastiaanssen, W. G. M., and Ali, S.: A new crop yield forecasting model based on satellite
835   measurements applied across the Indus Basin, Pakistan, Agriculture Ecosystems &
836   Environment, 94, 321-340, doi:10.1016/s0167-8809(02)00034-8, 2003.
837   Beckage, B., Gross, L. J., and Kauffman, S.: The limits to prediction in ecological systems,
838   Ecosphere, 2, doi:10.1890/es11-00211.1, 2011.
839   Bloom, A. A., Exbrayat, J. F., van der Velde, I. R., Feng, L., and Williams, M.: The decadal state
840   of the terrestrial carbon cycle: Global retrievals of terrestrial carbon allocation, pools, and
841   residence times, Proceedings of the National Academy of Sciences of the United States of
842   America, 113, 1285-1290, doi:10.1073/pnas.1515160113, 2016.
843   Botkin, D. B., Saxe, H., Araujo, M. B., Betts, R., Bradshaw, R. H. W., Cedhagen, T., Chesson, P.,
844   Dawson, T. P., Etterson, J. R., Faith, D. P., Ferrier, S., Guisan, A., Hansen, A. S., Hilbert, D. W.,
845   Loehle, C., Margules, C., New, M., Sobel, M. J., and Stockwell, D. R. B.: Forecasting the effects
846   of global warming on biodiversity, Bioscience, 57, 227-236, doi:10.1641/b570306, 2007.
847   Braswell, B. H., Sacks, W. J., Linder, E., and Schimel, D. S.: Estimating diurnal to annual
848   ecosystem parameters by synthesis of a carbon flux model with eddy covariance net
849   ecosystem exchange observations, Global Change Biology, 11, 335-355,
850   doi:10.1111/j.1365-2486.2005.00897.x, 2005.
851   Clark, J. S., Carpenter, S. R., Barber, M., Collins, S., Dobson, A., Foley, J. A., Lodge, D. M.,
852   Pascual, M., Pielke, R., Pizer, W., Pringle, C., Reid, W. V., Rose, K. A., Sala, O., Schlesinger, W.
853   H., Wall, D. H., and Wear, D.: Ecological forecasts: An emerging imperative, Science, 293,
854   657-660, doi:10.1126/science.293.5530.657, 2001.
855   Clark, J. S., Lewis, M., McLachlan, J. S., and HilleRisLambers, J.: Estimating population
856   spread: What can we forecast and how well?, Ecology, 84, 1979-1988, doi:10.1890/01-
857   0618, 2003.
858   Corbet, S. A., Saville, N. M., Fussell, M., PrysJones, O. E., and Unwin, D. M.: The competition
859   box: A graphical aid to forecasting pollinator performance, Journal of Applied Ecology, 32,
860   707-719, doi:10.2307/2404810, 1995.

861   Coreau, A., Pinay, G., Thompson, J. D., Cheptou, P. O., and Mermet, L.: The rise of research on
862   futures in ecology: rebalancing scenarios and predictions, Ecology Letters, 12, 1277-1286,
863   doi:10.1111/j.1461-0248.2009.01392.x, 2009.
864   Coreau, A., Treyer, S., Cheptou, P. O., Thompson, J. D., and Mermet, L.: Exploring the
865   difficulties of studying futures in ecology: what do ecological scientists think?, Oikos, 119,
866   1364-1376, doi:10.1111/j.1600-0706.2010.18195.x, 2010.
867   Craft, C., Clough, J., Ehman, J., Joye, S., Park, R., Pennings, S., Guo, H. Y., and Machmuller, M.:
868   Forecasting the effects of accelerated sea-level rise on tidal marsh ecosystem services,
869   Frontiers in Ecology and the Environment, 7, 73-78, doi:10.1890/070219, 2009.
870   Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K.: Accounting for uncertainty
871   in ecological analysis: the strengths and limitations of hierarchical statistical modeling,
872   Ecological Applications, 19, 553-570, doi:10.1890/07-0744.1, 2009.
873   Dietze, M. C., Lebauer, D. S., and Kooper, R.: On improving the communication between
874   models and data, Plant Cell and Environment, 36, 1575-1585, doi:10.1111/pce.12043,
875   2013.
876   Diez, J. M., Ibanez, I., Miller-Rushing, A. J., Mazer, S. J., Crimmins, T. M., Crimmins, M. A.,
877   Bertelsen, C. D., and Inouye, D. W.: Forecasting phenology: from species variability to
878   community patterns, Ecology Letters, 15, 545-553, doi:10.1111/j.1461-0248.2012.01765.x,
879   2012.
880   Ellison, A. M.: Repeatability and transparency in ecological research, Ecology, 91, 2536-
881   2539, doi:10.1890/09-0032.1, 2010.
882   Farquhar, G. D., Caemmerer, S. V., and Berry, J. A.: A biochemical-model of photosynthetic
883   CO_2 assimilation in leaves of C3 species, Planta, 149, 78-90, doi:10.1007/bf00386231,
884   1980.
885   Fordham, D. A., Akcakaya, H. R., Araujo, M. B., Elith, J., Keith, D. A., Pearson, R., Auld, T. D.,
886   Mellin, C., Morgan, J. W., Regan, T. J., Tozer, M., Watts, M. J., White, M., Wintle, B. A., Yates, C.,
887   and Brook, B. W.: Plant extinction risk under climate change: are forecast range shifts alone
888   a good indicator of species vulnerability to global warming?, Global Change Biology, 18,
889   1357-1371, doi:10.1111/j.1365-2486.2011.02614.x, 2012.
890   Gao, C., Wang, H., Weng, E. S., Lakshmivarahan, S., Zhang, Y. F., and Luo, Y. Q.: Assimilation of
891   multiple data sets with the ensemble Kalman filter to improve forecasts of forest carbon
892   dynamics, Ecological Applications, 21, 1461-1473, 2011.
893   Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L.,
894   Duke, C. S., and Porter, J. H.: Big data and the future of ecology, Frontiers in Ecology and the
895   Environment, 11, 156-162, doi:10.1890/120103, 2013.
896   Hanson, P. J., Riggs, J. S., Nettles, W. R., Phillips, J. R., Krassovski, M. B., Hook, L. A., Gu, L.,
897   Richardson, A. D., Aubrecht, D. M., Ricciuto, D. M., Warren, J. M., and Barbier, C.: Attaining
898   whole-ecosystem warming using air and deep-soil heating methods with an elevated CO_2
899   atmosphere, Biogeosciences, 14, 861-883, doi:10.5194/bg-14-861-2017, 2017.
900   Hare, J. A., Alexander, M. A., Fogarty, M. J., Williams, E. H., and Scott, J. D.: Forecasting the
901   dynamics of a coastal fishery species using a coupled climate-population model, Ecological
902   Applications, 20, 452-464, doi:10.1890/08-1863.1, 2010.
903   Huang, Y., Jiang, J., Ma, S., Ricciuto, D., Hanson, P. J., and Luo, Y.: Soil thermal dynamics, snow
904   cover and frozen depth under five temperature treatments in an ombrotrophic bog:
905   Constrained forecast with data assimilation, Journal of Geophysical Research:
906   Biogeosciences, doi:10.1002/2016JG003725, 2017.

907 Jiang, J., Huang, Y., Ma, S., Stacy, M., Shi, Z., Ricciuto, D. M., Hanson, P. J., and Luo, Y.:
908 Forecasting responses of a northern peatland carbon cycle to elevated CO2 and a gradient
909 of experimental warming, Journal of Geophysical Research: Biogeosciences,
910 doi:10.1002/2017jg004040, 2018.
911 Johnson, B. R., Kampe, T. U., and Kuester, M.: Development of airborne remote sensing
912 instrumentations for NEON, SPIE Optical Engineering+ Applications, 2010, 78090I-78090I-
913 78010,
914 Kearney, M. R., Wintle, B. A., and Porter, W. P.: Correlative and mechanistic models of
915 species distribution provide congruent forecasts under climate change, Conservation
916 Letters, 3, 203-213, doi:10.1111/j.1755-263X.2010.00097.x, 2010.
917 Keenan, T. F., Davidson, E., Moffat, A. M., Munger, W., and Richardson, A. D.: Using model-
918 data fusion to interpret past trends, and quantify uncertainties in future projections, of
919 terrestrial ecosystem carbon cycling, Global Change Biology, 18, 2555-2569,
920 doi:10.1111/j.1365-2486.2012.02684.x, 2012.
921 Kobori, H., Dickinson, J. L., Washitani, I., Sakurai, R., Amano, T., Komatsu, N., Kitamura, W.,
922 Takagawa, S., Koyama, K., Ogawara, T., and Miller-Rushing, A. J.: Citizen science: a new
923 approach to advance ecology, education, and conservation, Ecological Research, 31, 1-19,
924 doi:10.1007/s11284-015-1314-y, 2016.
925 Kolka, R. K., Sebestyen, S. D., Verry, E. S., and Brooks, K. N.: Peatland biogeochemistry and
926 watershed hydrology at the Marcell Experimental Forest, CRC Press Boca Raton 488 pp.,
927 2011.
928 Lebauer, D. S., Wang, D., Richter, K. T., Davidson, C. C., and Dietze, M. C.: Facilitating
929 feedbacks between field measurements and ecosystem models, Ecological Monographs, 83,
930 133-154, doi:10.1890/12-0137.1, 2013.
931 Liang, J. Y., Li, D. J., Shi, Z., Tiedje, J. M., Zhou, J. Z., Schuur, E. A. G., Konstantinidis, K. T., and
932 Luo, Y. Q.: Methods for estimating temperature sensitivity of soil organic matter based on
933 incubation data: A comparative evaluation, Soil Biology & Biochemistry, 80, 127-135,
934 doi:10.1016/j.soilbio.2014.10.005, 2015.
935 Ludascher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E. A., Tao, J., and
936 Zhao, Y.: Scientific workflow management and the Kepler system, Concurrency and
937 Computation-Practice & Experience, 18, 1039-1065, doi:10.1002/cpe.994, 2006.
938 Luo, Y. Q., and Reynolds, J. F.: Validity of extrapolating field CO2 experiments to predict
939 carbon sequestration in natural ecosystems, Ecology, 80, 1568-1583, doi:10.1890/0012-
940 9658(1999)080[1568:voefce]2.0.co;2, 1999.
941 Luo, Y. Q., Wan, S. Q., Hui, D. F., and Wallace, L. L.: Acclimatization of soil respiration to
942 warming in a tall grass prairie, Nature, 413, 622-625, doi:10.1038/35098065, 2001.
943 Luo, Y. Q., Melillo, J., Niu, S. L., Beier, C., Clark, J. S., Classen, A. T., Davidson, E., Dukes, J. S.,
944 Evans, R. D., Field, C. B., Czimczik, C. I., Keller, M., Kimball, B. A., Kueppers, L. M., Norby, R. J.,
945 Pelini, S. L., Pendall, E., Rastetter, E., Six, J., Smith, M., Tjoelker, M. G., and Torn, M. S.:
946 Coordinated approaches to quantify long-term ecosystem dynamics in response to global
947 change, Global Change Biology, 17, 843-854, doi:10.1111/j.1365-2486.2010.02265.x,
948 2011a.
949 Luo, Y. Q., Ogle, K., Tucker, C., Fei, S. F., Gao, C., LaDeau, S., Clark, J. S., and Schimel, D. S.:
950 Ecological forecasting and data assimilation in a data-rich era, Ecological Applications, 21,
951 1429-1442, 2011b.

952 Luo, Y. Q., Ahlstrom, A., Allison, S. D., Batjes, N. H., Brovkin, V., Carvalhais, N., Chappell, A.,
953 Ciais, P., Davidson, E. A., Finzi, A. C., Georgiou, K., Guenet, B., Hararuk, O., Harden, J. W., He, Y.
954 J., Hopkins, F., Jiang, L. F., Koven, C., Jackson, R. B., Jones, C. D., Lara, M. J., Liang, J. Y.,
955 McGuire, A. D., Parton, W., Peng, C. H., Randerson, J. T., Salazar, A., Sierra, C. A., Smith, M. J.,
956 Tian, H. Q., Todd-Brown, K. E. O., Torn, M., van Groenigen, K. J., Wang, Y. P., West, T. O., Wei,
957 Y. X., Wieder, W. R., Xia, J. Y., Xu, X., Xu, X. F., and Zhou, T.: Toward more realistic projections
958 of soil carbon dynamics by Earth system models, Global Biogeochemical Cycles, 30, 40-56,
959 doi:10.1002/2015gb005239, 2016.
960 Ma, S., Jiang, J., Huang, Y., Ricciuto, D., Hanson, P. J., and Luo, Y.: Data-constrained
961 projections of methane fluxes in a Northern Minnesota Peatland in response to elevated
962 CO2 and warming (Accepted), Journal of Geophysical Research: Biogeosciences, 2017.
963 Medlyn, B. E., Badeck, F. W., De Pury, D. G. G., Barton, C. V. M., Broadmeadow, M., Ceulemans,
964 R., De Angelis, P., Forstreuter, M., Jach, M. E., Kellomaki, S., Laitat, E., Marek, M., Philippot, S.,
965 Rey, A., Strassemeyer, J., Laitinen, K., Liozon, R., Portier, B., Roberntz, P., Wang, K., and
966 Jarvis, P. G.: Effects of elevated CO2 on photosynthesis in European forest species: a meta-
967 analysis of model parameters, Plant Cell and Environment, 22, 1475-1495,
968 doi:10.1046/j.1365-3040.1999.00523.x, 1999.
969 Melillo, J. M., Frey, S. D., DeAngelis, K. M., Werner, W. J., Bernard, M. J., Bowles, F. P., Pold, G.,
970 Knorr, M. A., and Grandy, A. S.: Long-term pattern and magnitude of soil carbon feedback to
971 the climate system in a warming world, Science, 358, 101-105,
972 doi:10.1126/science.aan2874, 2017.
973 Michener, W. K., and Jones, M. B.: Ecoinformatics: supporting ecology as a data-intensive
974 science, Trends in Ecology & Evolution, 27, 85-93, doi:10.1016/j.tree.2011.11.016, 2012.
975 Miller-Rushing, A., Primack, R., and Bonney, R.: The history of public participation in
976 ecological research, Frontiers in Ecology and the Environment, 10, 285-290,
977 doi:10.1890/110278, 2012.
978 Moorcroft, P. R.: How close are we to a predictive science of the biosphere?, Trends in
979 Ecology & Evolution, 21, 400-407, doi:10.1016/j.tree.2006.04.009, 2006.
980 Mouquet, N., Lagadeuc, Y., Devictor, V., Doyen, L., Duputie, A., Eveillard, D., Faure, D.,
981 Garnier, E., Gimenez, O., Huneman, P., Jabot, F., Jarne, P., Joly, D., Julliard, R., Kefi, S., Kergoat,
982 G. J., Lavorel, S., Le Gall, L., Meslin, L., Morand, S., Morin, X., Morlon, H., Pinay, G., Pradel, R.,
983 Schurr, F. M., Thuiller, W., and Loreau, M.: REVIEW: Predictive ecology in a changing world,
984 Journal of Applied Ecology, 52, 1293-1310, doi:10.1111/1365-2664.12482, 2015.
985 Niu, S. L., Luo, Y. Q., Dietze, M. C., Keenan, T. F., Shi, Z., Li, J. W., and Chapin, F. S.: The role of
986 data assimilation in predictive ecology, Ecosphere, 5, doi:10.1890/es13-00273.1, 2014.
987 Ong, J. B. S., Chen, M. I. C., Cook, A. R., Lee, H. C., Lee, V. J., Lin, R. T. P., Tambyah, P. A., and
988 Goh, L. G.: Real-Time Epidemic Monitoring and Forecasting of H1N1-2009 Using Influenza-
989 Like Illness from General Practice and Family Doctor Clinics in Singapore, Plos One, 5,
990 doi:10.1371/journal.pone.0010036, 2010.
991 Osterweil, L. J., Clarke, L. A., Ellison, A. M., Boose, E., Podorozhny, R., and Wise, A.: Clear and
992 Precise Specification of Ecological Data Management Processes and Dataset Provenance,
993 Ieee Transactions on Automation Science and Engineering, 7, 189-195,
994 doi:10.1109/tase.2009.2021774, 2010.
995 Parton, W. J., Stewart, J. W. B., and Cole, C. V.: Dynamics of c, n, p and s in grassland soils - a
996 model, Biogeochemistry, 5, 109-131, doi:10.1007/bf02180320, 1988.

997   Parton, W. J., Morgan, J. A., Wang, G. M., and Del Grosso, S.: Projected ecosystem impact of
998   the Prairie Heating and CO2 Enrichment experiment, New Phytologist, 174, 823-834,
999   doi:10.1111/j.1469-8137.2007.02052.x, 2007.
1000   Perretti, C. T., Munch, S. B., and Sugihara, G.: Model-free forecasting outperforms the correct
1001   mechanistic model for simulated and experimental data, Proceedings of the National
1002   Academy of Sciences of the United States of America, 110, 5253-5257,
1003   doi:10.1073/pnas.1216076110, 2013.
1004   Peylin, P., Bacour, C., MacBean, N., Leonard, S., Rayner, P., Kuppel, S., Koffi, E., Kane, A.,
1005   Maignan, F., Chevallier, F., Ciais, P., and Prunet, P.: A new stepwise carbon cycle data
1006   assimilation system using multiple data streams to constrain the simulated land surface
1007   carbon cycle, Geoscientific Model Development, 9, 3321-3346, doi:10.5194/gmd-9-3321-
1008   2016, 2016.
1009   Purves, D., Scharlemann, J., Harfoot, M., Newbold, T., Tittensor, D. P., Hutton, J., and Emmott,
1010   S.: Time to model all life on Earth, Nature, 493, 295-297, 2013.
1011   Robertson, G. P., Collins, S. L., Foster, D. R., Brokaw, N., Ducklow, H. W., Gragson, T. L., Gries,
1012   C., Hamilton, S. K., McGuire, A. D., and Moore, J. C.: Long-term ecological research in a
1013   human-dominated world, BioScience, 62, 342-353, 2012.
1014   Schaefer, K., Schwalm, C. R., Williams, C., Arain, M. A., Barr, A., Chen, J. M., Davis, K. J.,
1015   Dimitrov, D., Hilton, T. W., Hollinger, D. Y., Humphreys, E., Poulter, B., Raczka, B. M.,
1016   Richardson, A. D., Sahoo, A., Thornton, P., Vargas, R., Verbeeck, H., Anderson, R., Baker, I.,
1017   Black, T. A., Bolstad, P., Chen, J. Q., Curtis, P. S., Desai, A. R., Dietze, M., Dragoni, D., Gough, C.,
1018   Grant, R. F., Gu, L. H., Jain, A., Kucharik, C., Law, B., Liu, S. G., Lokipitiya, E., Margolis, H. A.,
1019   Matamala, R., McCaughey, J. H., Monson, R., Munger, J. W., Oechel, W., Peng, C. H., Price, D. T.,
1020   Ricciuto, D., Riley, W. J., Roulet, N., Tian, H. Q., Tonitto, C., Torn, M., Weng, E. S., and Zhou, X.
1021   L.: A model-data comparison of gross primary productivity: Results from the North
1022   American Carbon Program site synthesis, Journal of Geophysical Research-Biogeosciences,
1023   117, doi:10.1029/2012jg001960, 2012.
1024   Schimel, D.: The era of continental-scale ecology, Frontiers in Ecology and the Environment,
1025   9, 311-311, 2011.
1026   Schindler, D. E., and Hilborn, R.: Prediction, precaution, and policy under global change,
1027   Science, 347, 953-954, doi:10.1126/science.1261824, 2015.
1028   Scholze, M., Kaminski, T., Rayner, P., Knorr, W., and Giering, R.: Propagating uncertainty
1029   through prognostic carbon cycle data assimilation system simulations, Journal of
1030   Geophysical Research-Atmospheres, 112, doi:10.1029/2007jd008642, 2007.
1031   Shi, Z., Xu, X., Hararuk, O., Jiang, L. F., Xia, J. Y., Liang, J. Y., Li, D. J., and Luo, Y. Q.:
1032   Experimental warming altered rates of carbon processes, allocation, and carbon storage in
1033   a tallgrass prairie, Ecosphere, 6, doi:10.1890/es14-00335.1, 2015.
1034   Shi, Z., Yang, Y. H., Zhou, X. H., Weng, E. S., Finzi, A. C., and Luo, Y. Q.: Inverse analysis of
1035   coupled carbon-nitrogen cycles against multiple datasets at ambient and elevated CO2,
1036   Journal of Plant Ecology, 9, 285-295, doi:10.1093/jpe/rtv059, 2016.
1037   Sitch, S., Huntingford, C., Gedney, N., Levy, P. E., Lomas, M., Piao, S. L., Betts, R., Ciais, P., Cox,
1038   P., Friedlingstein, P., Jones, C. D., Prentice, I. C., and Woodward, F. I.: Evaluation of the
1039   terrestrial carbon cycle, future plant geography and climate-carbon cycle feedbacks using
1040   five Dynamic Global Vegetation Models (DGVMs), Global Change Biology, 14, 2015-2039,
1041   doi:10.1111/j.1365-2486.2008.01626.x, 2008.

1042 Steppe, K., von der Crone, J. S., and Pauw, D. J. W.: TreeWatch.net: A Water and Carbon
1043 Monitoring and Modeling Network to Assess Instant Tree Hydraulics and Carbon Status,
1044 Frontiers in Plant Science, 7, doi:10.3389/fpls.2016.00993, 2016.
1045 Stumpf, R. P., Tomlinson, M. C., Calkins, J. A., Kirkpatrick, B., Fisher, K., Nierenberg, K.,
1046 Currier, R., and Wynne, T. T.: Skill assessment for an operational algal bloom forecast
1047 system, Journal of Marine Systems, 76, 151-161, doi:10.1016/j.jmarsys.2008.05.016, 2009.
1048 Sugihara, G., May, R., Ye, H., Hsieh, C. H., Deyle, E., Fogarty, M., and Munch, S.: Detecting
1049 Causality in Complex Ecosystems, Science, 338, 496-500, doi:10.1126/science.1227079,
1050 2012.
1051 Thomas, R. Q., Brooks, E. B., Jersild, A. L., Ward, E., Wynne, R. H., Albaugh, T. J., Dinon-
1052 Aldridge, H., Burkhart, H. E., Domec, J., Fox, T. R., Gonzalez-Benecke, C. A., Martin, T. A.,
1053 Noormets, A., Sampson, D. A., and Teskey, R. O.: Leveraging 35 years of Pinus taeda
1054 research in the southeastern US to constrain forest carbon cycle predictions: regional data
1055 assimilation using ecosystem experiments, Biogeosciences, 14, 3525-3547, 2017.
1056 Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C. J. A., and Buytaert, W.: Web technologies for
1057 environmental Big Data, Environmental Modelling & Software, 63, 185-198,
1058 doi:10.1016/j.envsoft.2014.10.007, 2015.
1059 Walker, A. P., Hanson, P. J., De Kauwe, M. G., Medlyn, B. E., Zaehle, S., Asao, S., Dietze, M.,
1060 Hickler, T., Huntingford, C., Iversen, C. M., Jain, A., Lomas, M., Luo, Y. Q., McCarthy, H., Parton,
1061 W. J., Prentice, I. C., Thornton, P. E., Wang, S. S., Wang, Y. P., Warlind, D., Weng, E. S., Warren,
1062 J. M., Woodward, F. I., Oren, R., and Norby, R. J.: Comprehensive ecosystem model-data
1063 synthesis using multiple data sets at two temperate forest free-air CO2 enrichment
1064 experiments: Model performance at ambient CO2 concentration, Journal of Geophysical
1065 Research-Biogeosciences, 119, 937-964, doi:10.1002/2013jg002553, 2014.
1066 Wang, Y. P., and Leuning, R.: A two-leaf model for canopy conductance, photosynthesis and
1067 partitioning of available energy I: Model description and comparison with a multi-layered
1068 model, Agricultural and Forest Meteorology, 91, 89-111, doi:10.1016/s0168-
1069 1923(98)00061-6, 1998.
1070 Wang, Y. P., Law, R. M., and Pak, B.: A global model of carbon, nitrogen and phosphorus
1071 cycles for the terrestrial biosphere, Biogeosciences, 7, 2261-2282, doi:10.5194/bg-7-2261-
1072 2010, 2010.
1073 Ward, E. J., Holmes, E. E., Thorson, J. T., and Collen, B.: Complexity is costly: a meta-analysis
1074 of parametric and non-parametric methods for short-term population forecasting, Oikos,
1075 123, 652-661, doi:10.1111/j.1600-0706.2014.00916.x, 2014.
1076 Weng, E. S., and Luo, Y. Q.: Soil hydrological properties regulate grassland ecosystem
1077 responses to multifactor global change: A modeling analysis, Journal of Geophysical
1078 Research-Biogeosciences, 113, doi:10.1029/2007jg000539, 2008.
1079 Weng, E. S., and Luo, Y. Q.: Relative information contributions of model vs. data to short-
1080 and long-term forecasts of forest carbon dynamics, Ecological Applications, 21, 1490-1505,
1081 2011.
1082 Weng, E. S., Luo, Y. Q., Gao, C., and Oren, R.: Uncertainty analysis of forest carbon sink
1083 forecast with varying measurement errors: a data assimilation approach, Journal of Plant
1084 Ecology, 4, 178-191, doi:10.1093/jpe/rtr018, 2011.
1085 Wieder, W. R., Bonan, G. B., and Allison, S. D.: Global soil carbon projections are improved
1086 by modelling microbial processes, Nature Climate Change, 3, 909-912,
1087 doi:10.1038/nclimate1951, 2013.

1088   Xu, T., White, L., Hui, D. F., and Luo, Y. Q.: Probabilistic inversion of a terrestrial ecosystem
1089   model: Analysis of uncertainty in parameter estimation and model prediction, Global
1090   Biogeochemical Cycles, 20, doi:10.1029/2005gb002468, 2006.
1091   Zhou, T., and Luo, Y. Q.: Spatial patterns of ecosystem carbon residence time and NPP-
1092   driven carbon uptake in the conterminous United States, Global Biogeochemical Cycles, 22,
1093   doi:10.1029/2007gb002939, 2008.
1094   Zhou, X. H., Zhou, T., and Luo, Y. Q.: Uncertainties in carbon residence time and NPP-driven
1095   carbon uptake in terrestrial ecosystems of the conterminous USA: a Bayesian approach,
1096   Tellus Series B-Chemical and Physical Meteorology, 64, doi:10.3402/tellusb.v64i0.17223,
1097   2012.

1098

1099 **Tables**

1100 Table 1. Discussion stimulated by EcoPAD-SPRUCE forecasting among modellers and

1101 experimenters on how to improve predictions of the relative contribution of different pathways

1102 of methane emissions

	Discussion
1	No strong bubbles are noted at field and a non-observation constrained modelling study at a similar site from another project concluded minor ebullition contribution, which are at odds with TECO result.
2	$CH_4:CO_2$ ratio might explain the discrepancy. The other modelling study assumed that decomposed C is mainly turned into CO_2 and a smaller fraction is turned into CH_4. The large $CH_4:CO_2$ ratio at this site may result in higher CH4 flux. It seems that the most "flexible" term is ebullition because any "excess" (above saturation) CH4 is immediately released to ebullition, while the plant transport term is constrained by vegetation data.
3	Experimental researches on the relative contribution to methane emission from three different pathways are rare.
4	Current available observations include net surface flux of methane from the large collars, incubation data that should represent methane sources within the profile, and gas/DOC profile data that can indicate active zones within the peat profile. What are additional data needed to constrain relative contribution of different pathways?
5	I had always thought that peatlands don't bubble much, but the super-sensitive GPS measurements found movements of the surface of the GLAP peatlands consistent with degassing events, and subsurface radar images did show layers that were interpreted as bubble-layers.
6	Pore water gas data, perhaps N_2 or Ar may shed some light on the relative importance of ebullition.
7	It is really hard to accurately distinguish the three pathways. It has to rely on multiple approaches. Particularly for the SPRUCE site, the vegetation cover varies, vegetation species varies. How many channels each species has affect the transport? Meanwhile, the presence of plant (even not vascular plant) will lead to more gas transport, but as bubbles, rather than plant-mediated transport.
8	It depends on model structure and algorithm to simulate diffusion, vascular, and ebullition. Most models assume a threshold to allow ebullition. Diffusion is treated in similar ways as ebullition in some models (most one layer or two layers models). For the multiple layers models, the diffusion occurs from bottom to top mm by mm, layer by layer, therefore, the gas diffusion from top layer to atmosphere is considered the diffusion flux. If that is the case, the time step and wind speed and pressure matter (most models do not consider wind and pressure impacts). Plant transport is really dependent on the parameter for plant species, aerenchyma, etc. The gas transportability of plant is associated with biomass, NPP, or root biomass, seasonality of plant growth, etc. in models. All these differences might cause biases in the final flux.
9	With only the CH_4 emission data cannot constrain the relative contribution of three pathways. Concentration data in different soil layers may help constrain.
10	Diffusion coefficient calculation in TECO adopts the "three-porosity-model" which is ideal for mineral soil, but may not fit the organic soil. "Millington-Quirk model" for should be a better choice for peat soil.
11	The boundary condition should be taken care of, but it brings in more uncertainties including the wind speed and piston velocity, etc.,
12	CH_4 emissions captured in static chambers does not include the episodic ebullition events. So (1) the static chambers underestimate the total methane emission and (2) might need to exclude the ebullition pathway when using the observation data to constrain the CH_4 emission. But this point seems haven't been paid attention to in other models.

1103

1104 **Figure Legends**

1105 **Figure 1** Schema of approaches to forecast future ecological responses from common practice

1106 (the upper panel) and the Ecological Platform for Assimilation of Data (EcoPAD (v1.0)) (bottom

1107 panel). The common practice makes use of observations to develop or calibrate models to make

1108 predictions while the EcoPAD (v1.0) approach advances the common practice through its fully

1109 interactive platform. EcoPAD (v1.0) consists of four major components: experiment/data, model,

1110 data assimilation and the scientific workflow (green arrows or lines). Data and model are

1111 iteratively integrated through its data assimilation systems to improve forecasting. And its near-

1112 real time forecasting results are shared among research groups through its web interface to guide

1113 new data collections. The scientific workflow enables web-based data transfer from sensors,

1114 model simulation, data assimilation, forecasting, result analysis, visualization and reporting,

1115 encouraging broad user-model interactions especially for the experimenters and the general

1116 public with limited background in modelling. Images from the SPRUCE field experiments

1117 (https://mnspruce.ornl.gov/) are used to represent data collection and the flowchart of TECO

1118 model is used to delegate ecological models.

1119 **Figure 2** The data assimilation system inside the Ecological Platform for Assimilation of Data

1120 (EcoPAD (v1.0)) towards better forecasting of terrestrial carbon dynamics

1121 **Figure 3** The scientific workflow of EcoPAD (v1.0). The workflow wraps ecological models

1122 and data assimilation algorithms with the docker containerization platform. Users trigger

1123 different tasks through the Representational State Transfer (i.e., RESTful) application-

1124 programming interface (API). Tasks are managed through the asynchronous task queue, Celery.

1125 Tasks can be executed concurrently on a single or more worker servers across different scalable

1126   IT infrastructures. MongoDB is a database software that takes charge of data management in

1127   EcoPAD (v1.0) and RabbitMQ is a message broker.

1128

1129   **Figure 4.** Schema of interactive communication between modellers and experimenters through

1130   the prediction-question-discussion-adjustment-prediction cycle to improve ecological

1131   forecasting. The schema is inspired by an episode of experimenter-modeller communication

1132   stimulated by the EcoPAD-SPRUCE platform. The initial methane model constrained by static

1133   chamber methane measurements was used to predict relative contributions of three methane

1134   emission pathways (i.e., ebullition, plant mediated transportation (PMT) and diffusion) to the

1135   overall methane fluxes under different warming treatments (+ 0 °C, +2.25 °C, +4.5 °C, +6.75 °C

1136   and +9 °C). The initial results indicated a dominant contribution from ebullition especially under

1137   +9 °C which was doubted by experimenters. The discrepancy stimulated communications

1138   between modellers and experimenters with detailed information listed in Table 1. After extensive

1139   discussion, the model structure was adjusted and field observations were re-evaluated. And a

1140   second round of forecasting yielded more reliable predictions.

1141   **Figure 5.** Posterior distribution of the ratio of $CH_4:CO_2$ (panel a) and the temperature sensitivity

1142   of methane production (Q_{10_CH4}, panel b) under 5 warming treatments.

1143   **Figure 6.** Updated vs. un-updated forecasting of gross primary production (GPP, panels a,c) and

1144   soil organic C content (SoilC, panels b,d). The upper panels show 3 series of forecasting with

1145   updated vs. stochastically generated weather forcing. "updated" means the real meteorology

1146   forcing monitored from field weather station. Cyan indicates forecasting with 100 stochastically

1147   generated weather forcing from January 2015 to December 2024 (S1); red corresponds to

1148   updated forecasting with two stages, that is, updating with measured weather forcing from

1149   January 2015 to July 2016 followed by forecasting with 100 stochastically generated weather

1150   forcing from August 2016 to December 2024 (S2); and blue shows updated forecasting with

1151   measured weather forcing from January 2015 to December 2016 followed by forecasting with

1152   100 stochastically generated weather forcing from January 2017 to December 2024 (S3). The

1153   bottom panels display mismatches between updated forecasting (S2,3) and the original un-

1154   updated forecasting (S1). Red displays the difference between S2 and S1 (S2-S1) and blue shows

1155   discrepancy between S3 and S1 (S3-S1). Dashed green lines indicate the start of forecasting with

1156   stochastically generated weather forcing. Note that the left 2 panels are plotted on yearly time-

1157   scale and the right 2 panels show results on monthly time-scale.

1158

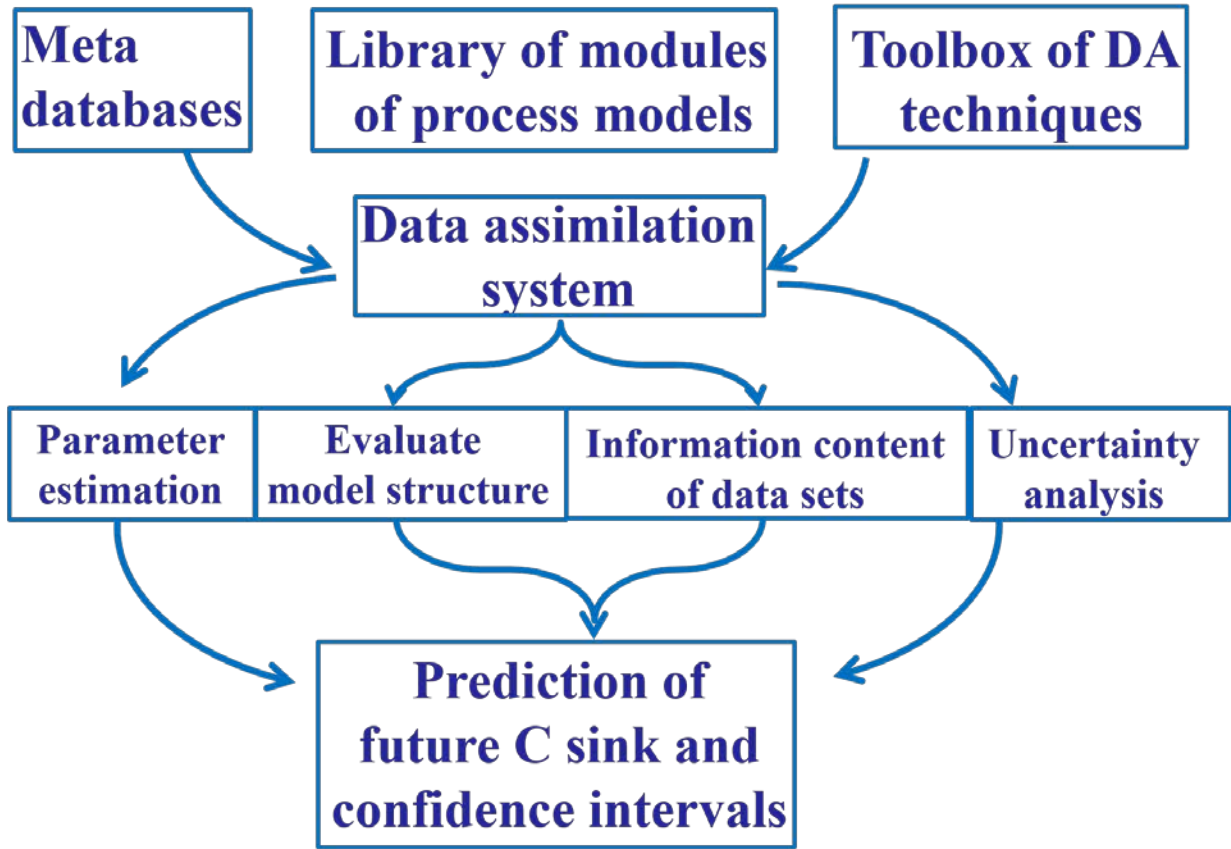1159

1160

1161 **Figure 1**

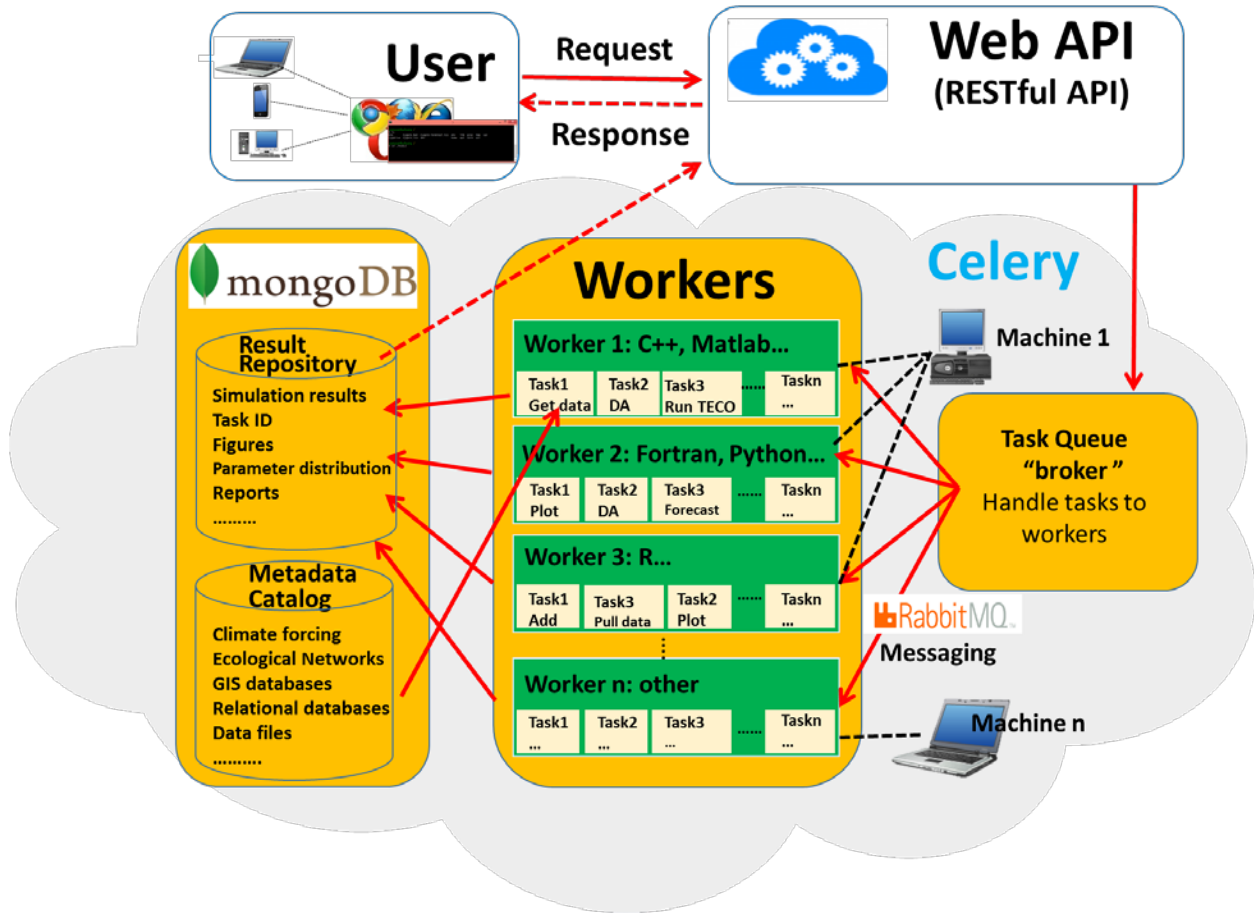1162
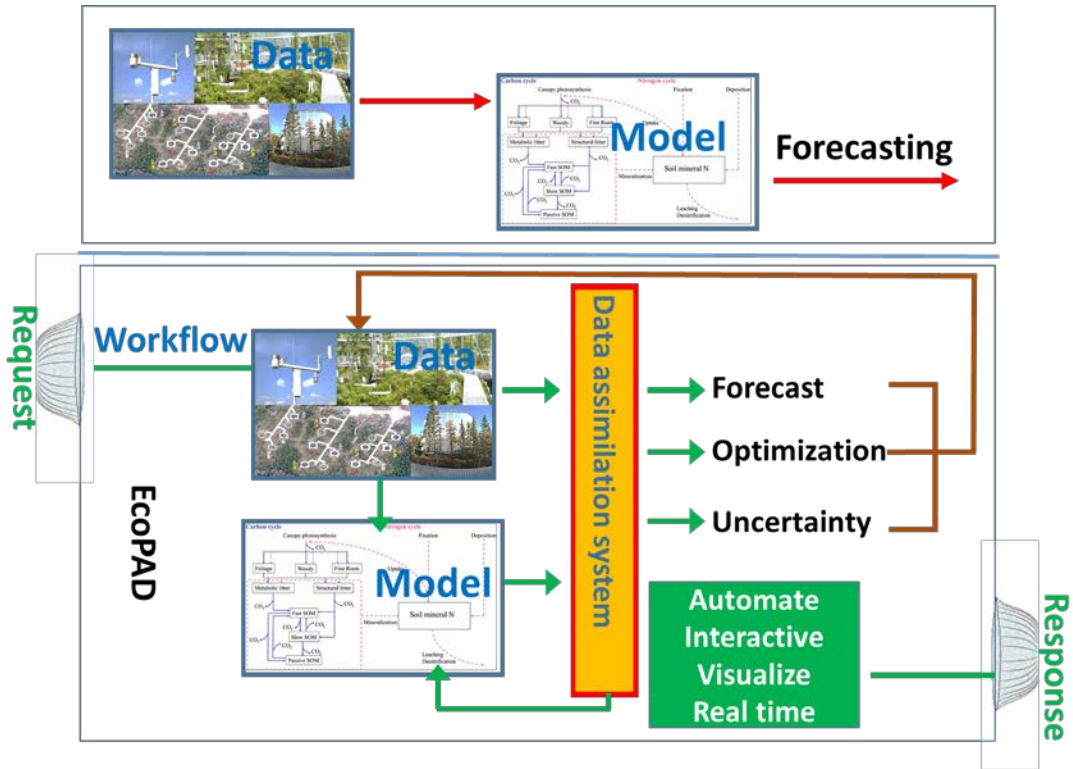
1163

1164

1165 **Figure 2**

1166

1167

1168

1169

1170

1171

1172   **Figure 3**

1173

1174

1175

50
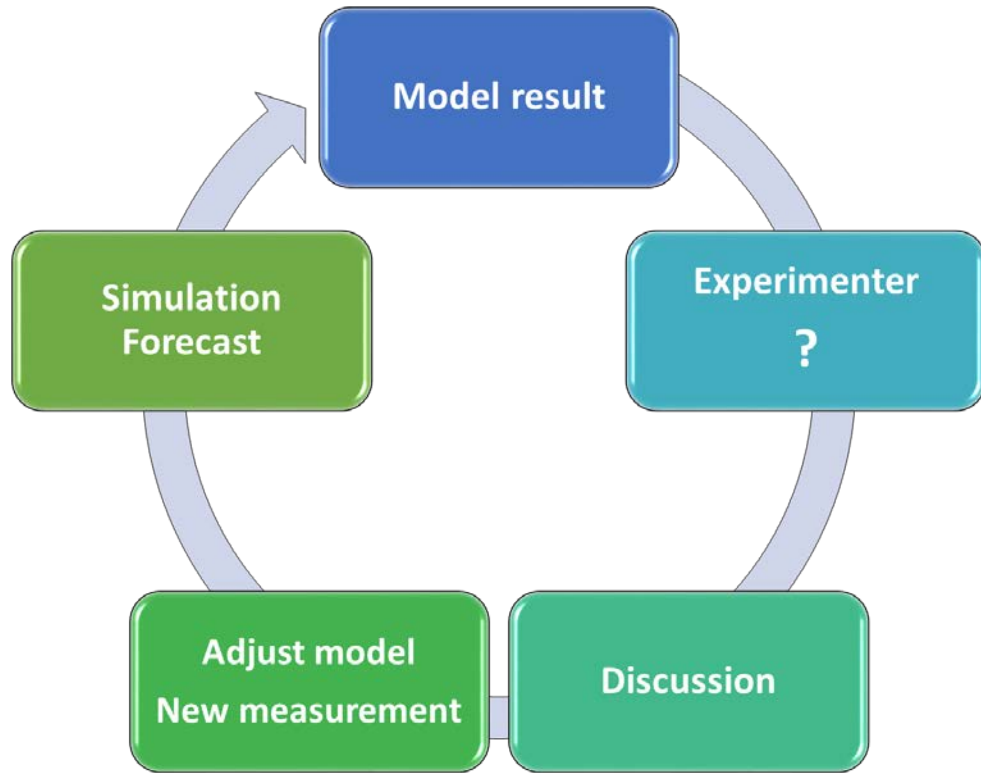
1176

1177   **Figure 4**
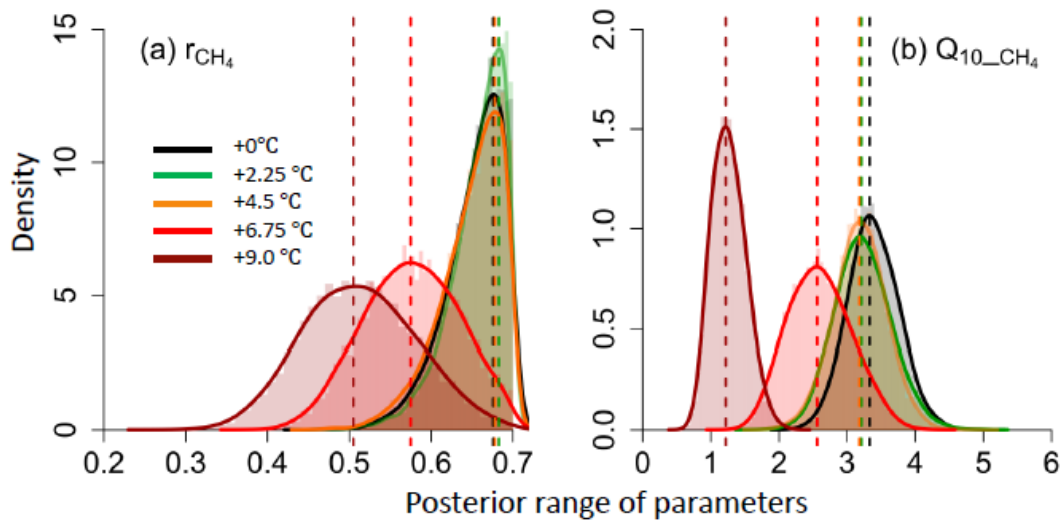
1178
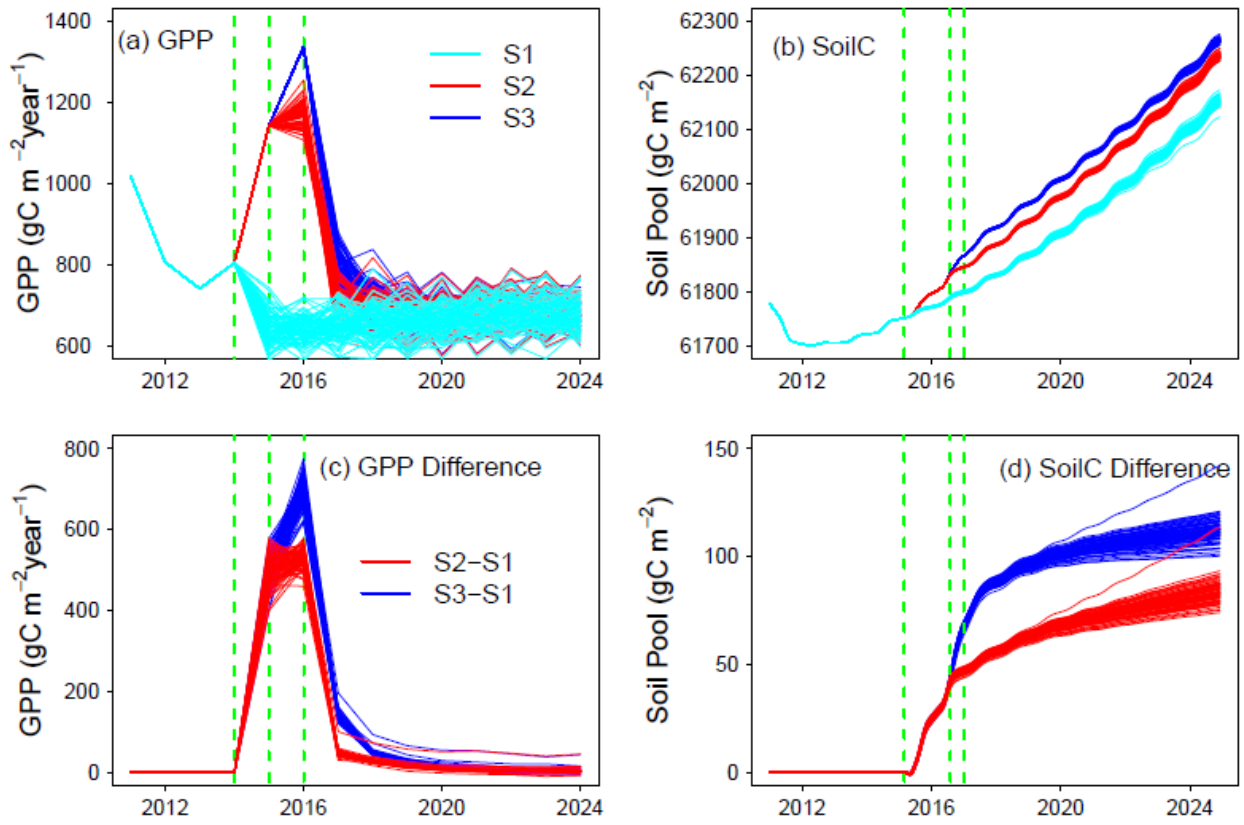

1179

1180

1181   **Figure 5**

1182

1183

1184

1185

1186

1187

1188   **Figure 6**

1189

1190