1

The role of data assimilation in predictive ecology

3

Shuli Niu[1,2], Yiqi Luo[2], Michael C. Dietze[3], Trevor F. Keenan[4,5], Zheng Shi[2], Jianwei Li[6,2], F. Stuart Chapin III[7]

6

[1]Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

[2]Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK 73072, USA

[3]Department of Earth and Environment, Boston University, Boston, MA 02215, USA

[4]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, 02138, USA

[5]Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109, Australia

[6]Institute of Agricultural Resources and Regional Planning Chinese Academy of Agricultural Sciences No. 12 ZhongGuanCun Southern Street, Beijing, 100081, P. R. China.

[7]Institute of Arctic Biology, University of Alaska Fairbanks, Fairbanks, AK 99775, USA

Abstract: In this rapidly changing world, improving the capacity to predict future dynamics of

ecological systems and their services is essential for better stewardship of the earth system.

Prediction relies on models that describe our understanding of the major processes that underlie

system dynamics and data about these processes and the present state of ecosystems. Prediction

becomes more effective when models are well informed by data. A technological revolution in

the capacity to collect data now provides very different opportunities to test hypotheses and

project future dynamics than when many standard statistical tests were first developed. Data

assimilation is an emerging statistical approach to combine models with data in a rigorous way to

constrain model parameters and system states, identify model error, and improve ecological

prediction. In this paper, we illustrate how data assimilation can improve ecological prediction to

support decision-making by reviewing applications of data assimilation across four different

research fields: (1) emerging infectious disease, (2) fisheries, (3) fire, and (4) the terrestrial

carbon cycle. Across these fields, data assimilation substantially improves prediction accuracy,

highlighting its important role in enabling predictive ecology. Data assimilation with regional

and global models faces major challenges, such as the large number of parameters to be

estimated, high computational demands, the need to integrate multiple and heterogeneous data

sets, and complex social-ecological interactions. Nevertheless, data assimilation provides an

important statistical approach that has great potential to enhance the predictive capacity of

ecological models in a changing climate.

Introduction

We live in a period marked by numerous environmental challenges: rapid climate change, profound alteration of biogeochemical cycles, unsustainable depletion of natural resources, rapid spread of invasive species, emerging infectious disease, and unprecedented threats from natural and anthropogenic disturbance (IPCC 2007). Predicting future changes in natural resources and the environment is therefore critical to effectively inform policy-makers if society hopes to continue to extract and use natural resources to support local-to-global economies and sustain thriving human societies. Toward that end, we, as a research community, need to advance the predictive capacity of ecology to better anticipate future states and services of ecosystems and enable a better stewardship of the earth system.

Despite the great societal need to predict future changes in the environment and natural resources under global change, predictive ecology has not been well developed, partly because we still lack effective approaches to improve the accuracy of ecological prediction. Data assimilation (DA) has the potential to enable and empower predictive ecology (Luo et al. 2011). DA refers to a suite of statistical techniques used to improve process models based on data. A process model describes how a system works by characterizing major components and their interactions under various forcing scenarios. Like regression analysis, DA requires both a model and data to be combined, aims to minimize differences between data and model, and uses an algorithm to obtain optimal parameters with data-model differences being minimized (Lewis et al. (2006). Unlike simple regression, DA can be applied to complex process models and multiple heterogeneous data sets, can optimizes tens or hundreds of parameters and state variables simultaneously, and has the capacity to navigate complex parameter spaces. As such, a DA-trained process model can not only better describe observed dynamics of an ecological system

but can also provide improved predictions of future states of the system in a manner consistent

with process understanding (Luo et al. 2011, Keenan et al. 2012). For example, it has long been

used to successfully improve numerical weather prediction. The capacity of DA to improve

ecological predictions, however, has not been fully explored.

This paper examines how DA may enable and empower predictive ecology. We first

discuss approaches used to predict future states of ecological systems. Then we use examples

from four research areas, infectious diseases, fisheries, fire, and the carbon cycle, to show that

DA can make predictive ecology possible by improving ecological predictions to the extent that

prediction results can guide policy making. We conclude with a discussion of the future

challenges and opportunities for further development of DA to enable predictive ecology.

Approaches of predicting future states of ecological systems

Ecological prediction involves describing future state of ecological systems with fully specified

uncertainties (Clark *et al*. 2001). There are two types of ecological prediction: 1) classical

prediction of the most likely future state of an ecological system, conditioned on current

observations and trends; and 2) projections, which are "what-if" analyses of the most likely

future state of a system under explicit scenarios of climate, land use, human population,

technologies, and economic activity (Luo *et al*. 2011). Classic prediction may be applied to fast-

evolving systems, such as the spread of an infectious disease, whose dynamics are strongly

governed by their current state, whereas projection becomes necessary when alternative

scenarios, such as disturbance impacts on ecosystem carbon dynamics, are plausible. Both types

of predictions must quantify the past and current states of ecological systems as a starting point

and then use models to project the future dynamics.

88   Traditionally, models have been the primary tool for predicting the future states of

89   ecological systems. For example, biogeochemical models have been incorporated into earth

90   system models to predict responses and feedbacks of the terrestrial carbon cycle under climate-

91   change scenarios (Lawrence and Swenson 2011). Those predictions have been incorporated into

92   the assessment reports of the Intergovernmental Panel on Climate Change (IPCC) to guide

93   mitigation and adaptation efforts by governments and public (IPCC 2007, 2013). The models

94   used have incorporated the dominant processes of ecological systems in order to quantitatively

95   explore ecosystem dynamics. Each of these models usually projects one deterministic trajectory

96   about the future behavior of the system. The trajectory, in most cases, does not capture the

97   dynamics of the ecosystem in the real world because using fixed values for all the parameters in

98   one deterministic model does not account for uncertainty in the parameters and state variables.

99   Therefore, models alone often do not represent the past and current system dynamics closely

100   enough to allow confidence in their predictions (Schwalm et al. 2010, Dietze et al. 2011, Keenan

101   et al. 2012).

102   To be useful for predictive ecology, we need both process models, to represent key

103   processes that determine the dynamic behavior of an ecological system, and also data, to identify

104   those key processes and constrain model parameters and state variables via data assimilation. DA

105   treats the model structure and ranges of parameter values as prior information in a Bayesian

106   framework to represent the current state of knowledge. It uses global optimization techniques to

107   update parameters and state variables of a model based on information contained in multiple,

108   heterogeneous data sets that describe the past and current states of an ecosystem (Fig. 1). The

109   posterior distributions of estimated parameters through DA usually include the maximum

110   likelihood estimates and are used for forward modeling toward prediction. The probability

5

111   density function of predicted future states after DA usually has a narrower spread than that

112   without DA for the same model structure and priors (Weng and Luo 2011; Keenan et al. 2013).

113   In the case that data does not match model structure and/or priors, data sets should not be

114   assimilated into a model before the model structure and/or priors are reexamined and adjusted

115   (LeBauer et al. 2013).

116   DA techniques have long been successfully applied to improve the accuracy of weather

117   forecasts by initializing an atmospheric model with estimates of recent and current weather that

118   form a basis for weather forecasts for the next few days (Kalnay 2002). Numerical weather

119   forecasting was first attempted in the 1920s but did not produce realistic forecasts until the 1950s

120   when DA could be done by computer simulation. Since then weather forecasts have improved

121   steadily and are currently much more accurate. This great success results from advances in (1)

122   mathematical models to represent physical processes of atmosphere dynamics, (2) improved

123   observation technologies, and (3) DA to infuse observations into models to continuously update

124   predictions. In DA, both model and data uncertainty are fully specified, as the information

125   content of a prediction is inversely proportional to its uncertainty (Chatfield 1995, Clark et al.

126   2001, Liu and Gupta 2007). It is also important to understand which processes drive predictive

127   uncertainty (Weng and Luo 2011, Dietze et al. 2013, LeBauer et al. 2013). Thus, data-model

128   integration via DA acknowledges the critical importance of not just data quantity, but also the

129   uncertainties in both the model and the data, in providing a predictive understanding. Overall,

130   DA can improve ecological prediction by (1) providing estimates of parameters, initial values,

131   and state variables, (2) quantifying uncertainties with respect to those parameters, initial

132   conditions, and modeled states of an ecological system, (3) selecting alternative model

133   structures, and (4) providing a quantitative basis to evaluate sampling strategies for future

134 experiments and observations that will enable improvement to models and predictions (Luo et al.

135 2011).

136   This paper mentions a few DA techniques, which mostly are variants of the Kalman filter

137 (KF). This is a recursive algorithm for estimating initial conditions, parameters, and state

138 variables of a model from a series of heterogeneous, intermittent measurements (Kalman 1960).

139 The KF iteratively repeats two sequential steps: forecast and update. The forecast step evolves

140 the currently estimated ecosystem state forward in time using the model. The update step adjusts

141 target parameters by combining observations of the current state of a system with the results

142 from a model. The Ensemble Kalman Filter (EnKF) is a commonly-used and flexible variation

143 on the KF that uses Monte Carlo techniques to generate ensemble predictions for the forecast

144 step (Gao et al. 2011). This paper mentions a few DA techniques, which mostly are variants of

145 KF. For a detailed description of DA techniques and principles, please refer to classical books

146 by Kalnay (2002), Lewis et al. (2006), and Evensen (2007) or some of the published papers,

147 including those by Harrison (2011), Peng et al. (2011), Wang et al. (2009) and Williams et al.

148 (2009).

149

150 **Application of DA in predictive ecology**

151 As a relatively new technique to ecology, DA has only recently been applied to a breadth of

152 research issues. In this section, we review examples of its applications in four diverse fields of

153 ecological research. In each field, we briefly describe the system and research issues, the

154 process-based models used to describe those systems, and the effectiveness of DA in

155 constraining the models and their predictions.

156

157 *Infectious disease*

158 The study of epidemics has long recognized the importance of connecting observations with

159 theory. DA is a powerful tool for rigorously establishing these connections, allowing predictions

160 of the trajectory of infectious disease outbreaks – an excellent example of the importance of DA

161 to societally relevant predictive ecology. The dynamics of infectious diseases are classically

162 described by Susceptible-Infected-Removed (SIR) models, which predict threshold responses to

163 both population size and R_0, the ratio between contagion and recovery rate (Kermack and

164 McKendrick 1927). Given these theoretical thresholds, DA has historically been solely focused

165 on estimating the parameters of SIR models (LaDeau et al. 2011). However, one of the primary

166 challenges in disease prediction is that in the early stages of an outbreak the key parameters

167 almost always have to be estimated from very few data points. Even with well-studied human

168 diseases like influenza, key parameters still need to be estimated each year from limited

169 information (Hooten et al. 2010). Combining this difficulty with the fact that the onset of an

170 outbreak is stochastic, non-linear, and sensitive to initial conditions, accurate prediction of

171 outbreaks remains challenging, whether it is for an animal, plant, or human disease.

172 A growing body of DA literature goes beyond just parameter estimation to demonstrate the

173 potential for real-time estimation and prediction of both parameters and state variables (such as

174 the number of people currently infected). In real-time estimation, updates to forecasts are

175 generated automatically based on new observations as soon as those observations are made. For

176 example, the emergence of real-time data on human influenza indirectly from Google Flu Trends

177 has spawned multiple attempts at real-time assimilation (Dukic et al. 2012, Shaman and

178 Karspeck 2012). The real-time modeling during the 2001 foot-and-mouth epidemic in the UK

179 (Keeling et al. 2003, Tildesley et al. 2008) contributed substantially to decisions to restrict

8

180 animal movement and cull livestock populations. These actions have been credited with helping

181 to control the outbreak, and more generally successes like these have contributed to progress of

182 disease modeling toward a more predictive science (Tildesley et al. 2008).

183   The real-time daily forecast of the 2009 H1N1 outbreak in Singapore is an excellent

184 example of the use of DA to quickly create an operational forecast system (Fig.2, Ong et al.

185 2010). This largely grass-roots effort was pulled together within a month with no budget during

186 the early stages of the outbreak by researchers at local hospitals and universities who asked local

187 clinics to report cases of influenza-like illness (Ong et al. 2010). A Susceptible-Exposed-

188 Infectious-Removed model was updated daily using a sequential DA approach to evolve the

189 estimates of both the model parameters and the numbers of individuals in each state. The

190 parameters of the model are estimated within the Bayesian statistical paradigm in which semi-

191 informative prior distributions are assigned to parameters and incoming data incorporated via the

192 likelihood function to obtain a time series of posterior distributions for the parameters and

193 unobserved state space. By doing so, DA increased the confidence in parameter estimates over

194 time, while allowing parameters to change as the epidemic evolved. From these daily updated

195 states and parameters, probabilistic forecasts of the epidemic's progression were generated for

196 the coming month and forwarded to a public website. Initial forecasts were adversely affected by

197 uncertainty in the parameters, caused by the vagueness of the subjective prior distributions and

198 the scarcity of information from the data (Ong et al. 2010). So, early in the epidemic the model

199 tended to over-predict the magnitude of the outbreak. However, the accuracy improved through

200 time. The model correctly predicted the timing of the peak infection weeks ahead of time and

201 provided a remarkably accurate forecast of the declining phase of the epidemic. These forecasts

202 were publicized by the local media and are believed to have contributed to increased

203 transparency, improved risk communication and mitigation, and reduced panic.

204   While much of the above discussion focused on human diseases there is also a growing

205 body of examples of DA in plant, animal, and zoonotic disease systems. Outside the context of

206 real-time forecasting, Bayesian state-space models have been used for combined parameter and

207 state estimation in systems ranging from measles in sub-Saharan Africa (Ferrari et al. 2008), to

208 white pine blister rust in the greater Yellowstone ecosystem (Hatala et al. 2011), to chronic

209 wasting disease of North American mule deer (Farnsworth et al. 2006). More generally,

210 hierarchical models are being used in disease modeling to better partition different sources of

211 uncertainty, such as separating process variability from observation error (LaDeau et al. 2007).

212

213 *Fisheries*

214 Sustainable exploitation of marine resources, such as commercial fisheries and fish farming, is

215 becoming increasingly important for economic development. Given past fishery collapses,

216 decision-makers need information on current and potential future states of fish stocks. DA

217 techniques, such as the Kalman filter, were first applied to fisheries models in the 1990s

218 (Schnute 1991). Now, a wide array of DA techniques has been applied to fishery models that

219 generally describe population dynamics and spatial distributions of fish presence and absence.

220   Fish stock assessment models are designed to determine the effects of the fishery on fish

221 populations and usually include the demographic processes of birth, natural death, harvest,

222 growth, maturation, and movement. The assessment models can be constrained either by time

223 series of fish catch to infer current and target fish stock abundance and the maximum sustainable

224 yield, or by a time series of detailed fishery catch-at-age data, to reconstruct the virtual

225 abundance of each annual cohort that had been fished (Methot and Wetzel 2013). For example,

226 after using DA to integrate standardized catch-per-unit-of-effort (CPUE) data into stock

227 assessment, the prediction uncertainty was reduced for simulation of the annual variation of

228 Trevally (*Pseudocaranx dentex*) in the west coast of New Zealand (Maunder and Langley 2004).

229   The Kalman filter (KF) is the most common DA technique used to improve fish stock

230 predictions. Holt and Peterman (2004) analyzed 24 sockeye salmon stocks and compared the

231 mean square error (MSE) and percent bias of predicted recruitments (relative vs. observed)

232 estimated with and without KF. They found that DA lowered MSE for about 35% of the stocks

233 and had bias closer to zero. Gronnevik and Evensen (2001) used DA for fisheries modeling and

234 stock assessment. Three DA techniques, EnKF, ensemble smoother (uses of data backward in

235 time to improve the estimates at prior times), and ensemble Kalman smoother (an extension of

236 EnKF to improve the estimate at prior times with data which work better with nonlinear

237 dynamics), were applied to an age-structured population model using catch-at-age data for

238 Icelandic cod. All three estimates that used DA had lower variance of prediction than that with

239 no data assimilation (Fig. 3) (Gronnevik and Evensen (2001).

240   In addition to stock assessment models, species distribution modeling is widely used to

241 estimate the presence and absence of fish species using the geographical and environmental

242 characteristics of each fishing location. To better predict fish occurrence, some models, such as a

243 hierarchical Bayesian spatial model (Munoz et al. 2013) and generalized linear and additive

244 models (GLM and GAM) (Guisan et al. 2002) were improved by DA in terms of either model

245 structure or parameters. For example, Bayesian Kriging, which can incorporate parameter

246 uncertainty into the prediction process, was used to generate the maps of horse mackerel

247 occurrence probabilities that that incorporated spatial and parameter uncertainties, geographical

248 characteristic data, and chlorophyll-a concentration (Munoz et al. 2013). In classical

249 geostatistical predictions, the true range of uncertainties is underestimated because many

250 parameters are estimated through the statistical model, potentially leading to the overestimation

251 of predictive accuracy. By incorporating the geographical characteristic data and chlorophyll-a

252 concentration data into the model, the occurrence of the species in unsampled areas was

253 predicted with more accuracy and fully quantified uncertainty.

254

255 *Fire*

256 Fire shapes the structure, function, and biodiversity of many ecosystems, such as savannah,

257 heathland and forest. Most fire management relies on the prediction of fire frequency, severity,

258 and spread (McKenzie et al. 2000, Xue et al. 2012) to assist decision-making. In addition, fire is

259 a key determinant of the global carbon cycle (Bowman et al. 2009).

260 Simulation models that describe fire behavior are used to manage fire. For example, Linn and

261 Cunningham (2005) modeled the forward spread of grassland fires under ambient atmosphere

262 winds and different initial lengths of the fire lines. AIOLOS-F, a model of fire behavior

263 prediction was developed as a decision support tool (Croba et al. 1994). However, many of the

264 existing models were not developed for practical use and real-time predictions, but instead used

265 for academic research. Accurate fire prediction requires not only physical models but also

266 datasets of fuel and weather, both of which are quite variable in both time and space (Keane et al.

267 2001). Moreover, fire behavior is also highly non-linear and complex with interactions among

268 the combustion processes, the landscape, local atmospheric environment and vegetation

269 characteristics, as well as human aspects (Lavorel et al. 2007). The performances of fire model

270 are also sensitive to boundary conditions, which are usually unknown (Sullivan 2009).

271 Therefore, there are many challenges in predicting fire behavior with sufficient accuracy to

272 support the decision-making process.

273   DA can address some of the above-mentioned issues in fire modeling and improve the

274 accuracy of fire predictions. For example, Mandel et al. (2008) used an ensemble Kalman filter

275 (EnKF) to assimilate measured temperature and remaining fuel into a simple wildfire model in

276 order to effectively track the location of the fireline. They found that fire temperature simulation

277 was significantly improved with EnKF, and the trained model accurately tracked the measured

278 fireline correctly regardless of ignition location. This successful application showed that even

279 with a relatively simple fire model and significant errors in the initial conditions, DA can help

280 train the model to realistically predict fire behavior with high confidence.

281   Sequential Monte Carlo (SMC or particle filters) is another effective DA technique that has

282 been used for integrating observations into wildfire models. Wildfire models typically have

283 nonlinear and unstable behaviors. SMC methods are particularly useful for the wildfire models

284 because they use an ensemble-based approach and implement the Bayesian recursion algorithms

285 directly to highly nonlinear state-space models (Doucet et al. 2001). Xue et al. (2012)

286 dynamically assimilated ground temperature sensor data of a wildfire to improve predictions.

287 The wildfire model is a discrete event model called DEVS-FIRE. SMC uses Bayesian inference

288 and a stochastic sampling technique to recursively estimate the state of dynamic systems from

289 given observations. After assimilating ground temperature sensor data, the DA system improved

290 the accuracy of wildfire prediction. The application of DA to fire behavior modeling is currently

291 not very common. Nonetheless, DA is a promising tool to help accurately predict fire behavior.

292

293 *Terrestrial carbon cycle*

294   The development of DA approaches has been an active area of research in terrestrial carbon

295   cycle models. This application is stimulated partly by climate change research and partly by an

296   increasing availability of data. From very simple models with scores of parameters to highly

297   complex models with thousands of parameters, biogeochemical models have been widely used to

298   explore ecosystem responses and feedbacks to climate changes on century-to-millennium time

299   scales (McGuire et al. 2001, Friedlingstein et al. 2006) or to study interactions of multiple

300   global-change factors on land management and ecosystem services on decadal or shorter time

301   scales (Schroter et al. 2005, Schmid et al. 2006, Pretzsch et al. 2008). DA has shown promise in

302   improving carbon cycle predictions (Williams et al. 2005, Luo et al. 2011, Peng et al. 2011),

303   spurred by increasing amounts of experimental data available and active process-based model

304   development,

305   In the early stages of research, DA was applied primarily to relatively simple terrestrial

306   ecosystem models and quickly became a very effective tool to diagnose model structure (Keenan

307   et al. 2012), evaluate the usefulness of different datasets (Keenan et al. 2013), and quantify

308   relative contributions of model and data in constraining model parameters and for carbon cycle

309   model predictions (Weng and Luo 2011). Several DA studies have indicated that different

310   datasets and their error structures can influence the parameter estimates and model predictions

311   (Luo et al. 2003, Xu et al. 2006, Richardson et al. 2010, Weng et al. 2012). Keenan et al. (2013),

312   for example, found that many datasets have redundant information for constraining model

313   performance. Only five out of the seventeen available data streams were necessary to constrain

314   the model (Fig. 4). In particular, numerous studies have demonstrated that flux data, such as soil

315   respiration and net ecosystem exchange (NEE), do not contain sufficient information to constrain

316   pool-related parameters (Wu et al. 2009, Keenan et al. 2013). Weng and Luo (2011) evaluated

317 information contributed by model structure vs. data to short- and long-term prediction of forest

318 carbon dynamics. Measurements over ten years at a forest ecosystem primarily constrained fast,

319 upstream carbon pools (e.g., foliage and fine root) whereas model structure determined slow,

320 downstream pools (e.g., slow and passive soil organic matter). This suggests that both process

321 understanding to improve model structure and datasets to constrain model parameters are

322 important for long-term carbon predictions.

323 Application of DA to regional and global carbon cycle models is challenging largely due

324 to prohibitive computational demands. Instead, comprehensive global models are often calibrated

325 against data at multiple locations. For example, Kuppel et al. (2012) assimilated net CO_2 flux

326 (NEE) and latent heat flux (LE) measurements collected from 12 temperate deciduous broadleaf

327 forest sites into the ORCHIDEE model. Several studies have successfully assimilated regional

328 and global data sets into models (Barrett 2003. , Zhou et al. 2009, Zhou et al. 2012, Hararuk et al.

329 2014). For example, Hararuk and Luo (in revision) applied DA to constrain the Community Land

330 Model (CLM3.5) against a global soil organic carbon database. The constrained model explained

331 41% of the global variability in the observed SOC in comparison with the initial 13%.

332 While increasingly employed for model calibration, DA has not been widely used for

333 predictions. Gao et al. (2011) used EnKF to assimilate eight sets of data from Duke Forest during

334 1996 to 2004 into a terrestrial ecosystem model as a basis for predicting the daily carbon pools

335 (state variables) from 2004 to 2012 (Fig. 5). Uncertainties in predicted carbon sinks increased

336 over time for the long-term carbon pools but remained constant over time for the short-term

337 carbon pools. In addition, prediction of future carbon dynamics requires weather data in the

338 future time as forcing since ecological system dynamics are influenced by weather and climate.

339 Thus, climate models and emissions scenarios are required to generate future weather data that

340   are used to drive forecasting with an ecological model. Keenan et al. (2013) showed how

341   prediction uncertainties to 2100 were greatly reduced by the incorporation of orthogonal data

342   constraints. The identification of the key datasets that contain the information needed to improve

343   model predictions of future states and fluxes remains an important and outstanding challenge.

344

345   **Future challenges and opportunities of DA**

346   Overall, all of the four areas discussed here use models to represent major processes underlying

347   the system dynamics. When models are not constrained by data before they are used to project

348   future states, the uncertainty in those projections is typically quite large. When models are

349   constrained by data, uncertainty in predictions declines and accuracy increases. Among all four

350   areas, there are a few examples of real-time forecasting capability. In the case of infectious

351   diseases and wildfire, real-time forecast can save human lives and thus has immediate societal

352   value. Natural resource managers, such as in fisheries and agriculture, would likewise benefit

353   from frequent, on-demand access to forecast models that have assimilated the most up-to-date

354   information. Likewise, both in natural resources and in other sectors of the economy, there is

355   increasingly concern about the carbon cycle impacts of human activity. However, decision

356   makers rarely have access to the best available models or data when evaluating the sustainability

357   of alternative scenarios, especially those operating at a local scale. Nevertheless, DA may not

358   improve prediction when ecological processes are not well understood or never observed (Luo et

359   al. 2011), and so any projections will always be contingent on model structure.

360   While DA has the potential to advance predictive ecology, future DA applications have to

361   deal with several challenges, such as: combining multiple data sources, assimilating data across

362   diverse scales and process, complex model structures, models with large numbers of parameters

363 to be estimated, high computational demands, and the need to expand applications into complex

364 social-ecological systems.

365

366 *Models become more and more complex: challenges and opportunities*

367 One of the challenges to further advance predictive ecology is related to complex model

368 structures. Terrestrial carbon models are continuously being developed, driven by our evolving

369 understanding of terrestrial ecosystems. For example, recent advances have incorporated detailed

370 descriptions of nutrient dynamics (Zaehle et al. 2010), community dynamics (Medvigy et al.

371 2009), and responses to disturbances (Prentice et al. 2011) in attempts to simulate the carbon

372 cycle in the real world as realistically as possible. As a consequence, models become more

373 complex but less tractable. DA with complex global models has become infeasible, as most of

374 DA methods are computationally expensive, unless innovative methods are developed for DA or

375 model structure analysis.

376   In addition, complex models make it very difficult to separate model structural error from

377 parameter error when using traditional model-comparison approaches (Keenan et al. 2011).

378 Recent advances in theoretical analysis of model structure have the potential to better identify

379 model structural and parameter errors (Xia et al. 2013). Advances in the analysis of model

380 uncertainty are likewise allowing better identification of model parameter error (LeBauer et al.

381 2013, Dietze et al. 2013). Although DA is not as easily applied to complex models as a whole,

382 key components of models can be identified for targeted improvement.

383

384 *Data become more and more available: challenges and opportunities*

385 The field of ecology is rapidly becoming data-rich with new automated measurement devices,

386   wireless networking, and improved cyber infrastructure greatly facilitate the collection of large

387   amounts of data. Current initiatives in continental-scale monitoring through ''ecological

388   observatories'', such as NEON (http://www.neoninc.org), ICOS (http://www.icos-

389   infrastructure.eu), FLUXNET (http:// www.fluxnet.ornl.gov; http://www.fluxdata.org), and

390   LTER (http://www.lternet.edu), promise an even greater data deluge in the future. As scientists

391   deal with larger and larger datasets, advanced techniques, such as data mining (e.g., (Moffat et

392   al. 2010) and model uncertainty analysis (e.g., (Keenan et al. 2011), are becoming more widely

393   applied.

394   The integration of the newly available data with ecological models is a clear direction

395   forward. It has been shown that using multiple data streams reduces model uncertainty

396   disproportionate to simply the increase in sample size, both for present–day model performance

397   (Richardson et al. 2010) and for projecting ecosystem function under future climate change

398   (Keenan et al. 2012). That said, not all data types give an equivalent improvement in model

399   performance, and indeed some data streams are redundant given the availability of others

400   (Keenan et al. 2013). The true information content of high frequency vs. low frequency

401   observations has also yet to be systematically assessed (e.g., (Weng and Luo 2011). Despite

402   challenges, it becomes essential to use multiple constraints to inform continental (Haverd et al.

403   2013) and global (Smith et al. 2013) models.

404

405   *Ecological issues become more complex in coupled human-nature systems*

406   Many of the most important biophysical changes occurring today are substantially influenced by

407   human decisions and actions, such as changes in climate, land-cover, and species composition.

408   These ecological changes, in turn, strongly affect society and its management of natural

409 resources. Future predictions are therefore best made in a social-ecological context, recognizing

410 the important feedbacks between the human and ecological components of social-ecological

411 systems. Social scientists are generally reluctant to forecast human actions, and therefore the

412 biophysical consequences of these actions, because people respond not only to past events but

413 also make decisions in response to perceptions of potential future outcomes.

414 DA has been most successful in improving predictions of the responses of ecological

415 systems to well-understood causes of variation. Building on this experience, it is most likely to

416 be useful for predicting trajectories of rapidly changing variables that follow predictable

417 temporal patterns and also respond predictably to events that are difficult to anticipate. Tropical

418 deforestation, for example, responds predictably to road construction in combination with other

419 variables (Geist and Lambin 2002), but the institutional factors that precipitate road development

420 are less predictable. DA that tracks road construction may therefore be a useful predictor of the

421 locations and rates of deforestation and associated declines in carbon sequestration and

422 biodiversity. Stewardship, which involves shaping the future of social-ecological systems to

423 facilitate ecosystem resilience and human well-being (Chapin et al. 2011), would benefit from

424 the use of DA to track the trajectories of these systems toward or away from critical transitions

425 and thresholds.

426

427 *Real- or near-real-time predictions*

428 To make predictions accessible for policy-making, we need to develop real- or near-time

429 forecasting capability, i.e. use the most currently available data to get updated predictions. Real-

430 time forecast has been successfully applied to infectious diseases and should be valuable for fire

431 behavior to save human lives and thus have immediate societal value. Near-time predictions can

432   have great economic and societal importance for fisheries and carbon cycle, as their economic

433   values and societal benefits can also be realized in long-term predictive estimates of ecological

434   processes. To realize the real- or near-time prediction, we not only need informed datasets,

435   process-oriented models and DA but also cyber-infrastructure to support data assimilation (Luo

436   et al. 2011, Dietze et al. 2013).

437

445

446   **Reference**

447   Barrett, A. 2003. . National Operational Hydrologic Remote Sensing Center SNOw Data Assimilation System
448   (SNODAS) Products at NSIDC. NSIDC Special Report 11., Boulder, CO, USA, National Snow and Ice Data
449   Center. Digital media.
450   Bowman, D. M. J. S., J. K. Balch, P. Artaxo, W. J. Bond, J. M. Carlson, M. A. Cochrane, C. M. D'Antonio, R. S. DeFries,
451   J. C. Doyle, S. P. Harrison, F. H. Johnston, J. E. Keeley, M. A. Krawchuk, C. A. Kull, J. B. Marston, M. A.
452   Moritz, I. C. Prentice, C. I. Roos, A. C. Scott, T. W. Swetnam, G. R. van der Werf, and S. J. Pyne. 2009. Fire in
453   the Earth System. Science **324**:481-484.
454   Chapin, F. S., M. E. Power, and J. J. Cole. 2011. Coupled biogeochemical cycles and Earth stewardship. Frontiers in
455   Ecology and the Environment **9**:3-3.
456   Chatfield, C. 1995. Model uncertainty, data mining and statistical inference. Royal Society Statistical Society Series
457   A, **158**:419-466.
458   Clark, J. S., S. R. Carpenter, M. Barber, S. Collins, A. Dobson, J. A. Foley, D. M. Lodge, M. Pascual, R. Pielke, W. Pizer,
459   C. Pringle, W. V. Reid, K. A. Rose, O. Sala, W. H. Schlesinger, D. H. Wall, and D. Wear. 2001. Ecological
460   forecasts: An emerging imperative. Science **293**:657-660.
461   Croba, D., D. lalas, C. Papadopoulos, and D. Tryfonopoulos. 1994. Numerical simulation of forest fire propagation in
462   complex terrain. Pages 491-500 Procedding of the 2nd international conference on forest fire research,
463   Coimbra, Portugal.
464   Dietze, M. C., D. S. Lebauer, and R. Kooper. 2013. On improving the communication between models and data.
465   Plant Cell Environ **36**:1575-1585.

466   Dietze, M. C., R. Vargas, A. D. Richardson, P. C. Stoy, A. G. Barr, R. S. Anderson, M. A. Arain, I. T. Baker, T. A. Black, J.
467   M. Chen, P. Ciais, L. B. Flanagan, C. M. Gough, R. F. Grant, D. Hollinger, R. C. Izaurralde, C. J. Kucharik, P.
468   Lafleur, S. G. Liu, E. Lokupitiya, Y. Q. Luo, J. W. Munger, C. H. Peng, B. Poulter, D. T. Price, D. M. Ricciuto,
469   W. J. Riley, A. K. Sahoo, K. Schaefer, A. E. Suyker, H. Q. Tian, C. Tonitto, H. Verbeeck, S. B. Verma, W. F.
470   Wang, and E. S. Weng. 2011. Characterizing the performance of ecosystem models across time scales: A
471   spectral analysis of the North American Carbon Program site-level synthesis. Journal of Geophysical
472   Research-Biogeosciences **116**:DOI: 10.1029/2011JG001661.
473   Doucet, A., N. d. Freitas, and N. Gordon. 2001. Sequential Monte Carlo Methods in Practice. Springer, New York.
474   Evensen, G. 2007. Data Assimilation:The Ensemble Kalman Filter. Springer, New York.
475   Farnsworth, M. L., J. A. Hoeting, N. T. Hobbs, and M. W. Miller. 2006. Linking chronic wasting disease to mule deer
476   movement scales: A hierarchical Bayesian approach. Ecological Applications **16**:1026-1036.
477   Friedlingstein, P., P. Cox, R. Betts, L. Bopp, W. Von Bloh, V. Brovkin, P. Cadule, S. Doney, M. Eby, I. Fung, G. Bala, J.
478   John, C. Jones, F. Joos, T. Kato, M. Kawamiya, W. Knorr, K. Lindsay, H. D. Matthews, T. Raddatz, P. Rayner,
479   C. Reick, E. Roeckner, K. G. Schnitzler, R. Schnur, K. Strassmann, A. J. Weaver, C. Yoshikawa, and N. Zeng.
480   2006. Climate-carbon cycle feedback analysis: Results from the (CMIP)-M-4 model intercomparison.
481   Journal of Climate **19**:3337-3353.
482   Gao, C., H. Wang, E. S. Weng, S. Lakshmivarahan, Y. F. Zhang, and Y. Q. Luo. 2011. Assimilation of multiple data sets
483   with the ensemble Kalman filter to improve forecasts of forest carbon dynamics. Ecological Applications
484   **21**:1461-1473.
485   Geist, H. J. and E. F. Lambin. 2002. Proximate causes and underlying driving forces of tropical deforestation.
486   Bioscience **52**:143-150.
487   Gronnevik, R. and G. Evensen. 2001. Application of ensemble-based techniques in fish stock assessment. Sarsia
488   **86**:517-526.
489   Guisan, A., T. C. Edwards, and T. Hastie. 2002. Generalized linear and generalized additive models in studies of
490   species distributions: setting the scene. Ecological Modelling **157**:89-100.
491   Hararuk, O., J. Xia, and Y. Luo. 2014. Evaluation and Improvement of a Global Land Model against Soil Carbon Data
492   Using a Bayesian MCMC Method. Journal of Geophysical Research-Biogeosciences **119**:doi:
493   10.1002/2013JG002535
494   Harrison, F. 2011. Getting started with meta-analysis. Methods in Ecology and Evolution **2**:1-10.
495   Haverd, V., M. R. Raupach, P. R. Briggs, J. G. Canadell, P. Isaac, C. Pickett-Heaps, S. H. Roxburgh, E. van Gorsel, R. A.
496   V. Rossel, and Z. Wang. 2013. Multiple observation types reduce uncertainty in Australia's terrestrial
497   carbon and water cycles. Biogeosciences **10**:2011-2040.
498   Holt, C. A. and R. M. Peterman. 2004. Long-term trends in age-specific recruitment of sockeye salmon
499   (Oncorhynchus nerka) in a changing environment. Canadian Journal of Fisheries and Aquatic Sciences
500   **61**:2455-2470.
501   Hooten, M. B., J. Anderson, and L. A. Waller. 2010. Assessing North American influenza dynamics with a statistical
502   SIRS model. Spatial and Spatio-Temporal Epidemiology **1**:177-185.
503   IPCC. 2007. IPCC WGI Fourth Assessment Report. Geneva: Intergovernmental Panel on Climate Change.
504   IPCC. 2013. IPCC Fifth Assessment Report: Climate Change 2013 (AR5). Cambridge University Press, Cambridge,
505   United Kingdom and New York, NY, USA.
506   Kalman, R. E. 1960. A new approach to linear filtering and prediction problems. Journal of Basic Engineering **82**:35-
507   45.
508   Kalnay, E. 2002. Atmospheric modeling, data assimilation and predictability. Cambridge University press,
509   Cambridge, UK.
510   Keane, R. E., R. Burgan, and J. van Wagtendonk. 2001. Mapping wildland fuels for fire management across multiple
511   scales: Integrating remote sensing, GIS, and biophysical modeling. International Journal of Wildland Fire
512   **10**:301-319.
513   Keeling, M. J., M. E. J. Woolhouse, R. M. May, G. Davies, and B. T. Grenfell. 2003. Modelling vaccination strategies
514   against foot-and-mouth disease. Nature **421**:136-142.
515   Keenan, T. F., M. S. Carbone, M. Reichstein, and A. D. Richardson. 2011. The model-data fusion pitfall: assuming
516   certainty in an uncertain world. Oecologia **167**:587-597.

517 Keenan, T. F., E. Davidson, A. M. Moffat, W. Munger, and A. D. Richardson. 2012. Using model-data fusion to
518   interpret past trends, and quantify uncertainties in future projections, of terrestrial ecosystem carbon
519   cycling. Global Change Biology **18**:2555-2569.
520 Keenan, T. F., E. A. Davidson, J. W. Munger, and A. D. Richardson. 2013. Rate my data: quantifying the value of
521   ecological data for the development of models of the terrestrial carbon cycle. Ecological Applications
522   **23**:273-286.
523 Kermack, W. O. and A. G. McKendrick. 1927. A Contribution to the Mathematical Theory of Epidemics. Proceedings
524   of the Royal Society A: Mathematical, Physical and Engineering Sciences **115**:700-721.
525 Kuppel, S., P. Peylin, F. Chevallier, C. Bacour, F. Maignan, and A. D. Richardson. 2012. Constraining a global
526   ecosystem model with multi-site eddy-covariance data. Biogeosciences **9**:3757-3776.
527 LaDeau, S. L., G. E. Glass, N. T. Hobbs, A. Latimer, and R. S. Ostfeld. 2011. Data-model fusion to better understand
528   emerging pathogens and improve infectious disease forecasting. Ecological Applications **21**:1443-1460.
529 Lavorel, S., M. D. Flannigan, E. F. Lambin, and M. C. Scholes. 2007. Vulnerability of land systems to fire: interactions
530   among humans, climate, the atmosphere, and ecosystems. Mitigation and Adaptation Strategies for
531   Global Change **12**:33-53.
532 Lawrence, D. M. and S. C. Swenson. 2011. Permafrost response to increasing Arctic shrub abundance depends on
533   the relative influence of shrubs on local soil cooling versus large-scale climate warming. Environmental
534   Research Letters **6**:Doi 10.1088/1748-9326/1086/1084/045504.
535 LeBauer, D. S., D. Wang, K. T. Richter, C. C. Davidson, and M. C. Dietze. 2013. Facilitating feedbacks between field
536   measurements and ecosystem models. Ecological Monographs **83**:133-154.
537 Lewis, J. M., S. Lakshmivarahan, and S. Dhall. 2006. Dynamic Data Assimilation : A Least Squares Approach.
538   Cambridge University Press, Cambridge, United Kingdom.
539 Linn, R. R. and P. Cunningham. 2005. Numerical simulations of grass fires using a coupled atmosphere-fire model:
540   Basic fire behavior and dependence on wind speed. Journal of Geophysical Research-Atmospheres
541   **110**:Doi 10.1029/2004jd005597.
542 Liu, Y. Q. and H. V. Gupta. 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation
543   framework. Water Resources Research **43**:Doi 10.1029/2006wr005756.
544 Luo, Y. Q., K. Ogle, C. Tucker, S. F. Fei, C. Gao, S. LaDeau, J. S. Clark, and D. S. Schimel. 2011. Ecological forecasting
545   and data assimilation in a data-rich era. Ecological Applications **21**:1429-1442.
546 Luo, Y. Q., L. W. White, J. G. Canadell, E. H. DeLucia, D. S. Ellsworth, A. C. Finzi, J. Lichter, and W. H. Schlesinger.
547   2003. Sustainability of terrestrial carbon sequestration: A case study in Duke Forest with inversion
548   approach. Global Biogeochemical Cycles **17**:Doi 10.1029/2002gb001923.
549 Mandel, J., L. S. Bennethum, J. D. Beezley, J. L. Coen, C. C. Douglas, M. Kim, and A. Vodacek. 2008. A wildland fire
550   model with data assimilation. Mathematics and Computers in Simulation **79**:584-606.
551 Maunder, M. N. and A. D. Langley. 2004. Integrating the standardization of catch-per-unit-of-effort into stock
552   assessment models: testing a population dynamics model and using multiple data types. Fisheries
553   Research **70**:389-395.
554 McGuire, A. D., S. Sitch, J. S. Clein, R. Dargaville, G. Esser, J. Foley, M. Heimann, F. Joos, J. Kaplan, D. W. Kicklighter,
555   R. A. Meier, J. M. Melillo, B. Moore, I. C. Prentice, N. Ramankutty, T. Reichenau, A. Schloss, H. Tian, L. J.
556   Williams, and U. Wittenberg. 2001. Carbon balance of the terrestrial biosphere in the twentieth century:
557   Analyses of CO2, climate and land use effects with four process-based ecosystem models. Global
558   Biogeochemical Cycles **15**:183-206.
559 McKenzie, D., D. L. Peterson, and J. K. Agee. 2000. Fire frequency in the interior Columbia River Basin: Building
560   regional models from fire history data. Ecological Applications **10**:1497-1516.
561 Medvigy, D., S. C. Wofsy, J. W. Munger, D. Y. Hollinger, and P. R. Moorcroft. 2009. Mechanistic scaling of
562   ecosystem function and dynamics in space and time: Ecosystem Demography model version 2. Journal of
563   Geophysical Research-Biogeosciences **114**:Doi 10.1029/2008jg000812.
564 Methot, R. D. and C. R. Wetzel. 2013. Stock synthesis: A biological and statistical framework for fish stock
565   assessment and fishery management. Fisheries Research **142**:86-99.
566 Moffat, A. M., C. Beckstein, G. Churkina, M. Mund, and M. Heimann. 2010. Characterization of ecosystem
567   responses to climatic controls using artificial neural networks. Global Change Biology **16**:2737-2749.

568 Munoz, F., M. G. Pennino, D. Conesa, A. Lo´pez-Qui´lez, and J. M. Bellido. 2013. Estimation and prediction of the
569   spatial occurrence of fish species using Bayesian latent Gaussian models. Stoch Environment Research
570   Risk Assessment **27**:1171-1180.
571 Ong, J. B. S., M. I. C. Chen, A. R. Cook, H. C. Lee, V. J. Lee, R. T. P. Lin, P. A. Tambyah, and L. G. Goh. 2010. Real-Time
572   Epidemic Monitoring and Forecasting of H1N1-2009 Using Influenza-Like Illness from General Practice and
573   Family Doctor Clinics in Singapore. Plos One **5**:DOI 10.1371/journal.pone.0010036.
574 Peng, C. H., J. Guiot, H. B. Wu, H. Jiang, and Y. Q. Luo. 2011. Integrating models with data in ecology and
575   palaeoecology: advances towards a model-data fusion approach. Ecology Letters **14**:522-536.
576 Prentice, I. C., D. I. Kelley, P. N. Foster, P. Friedlingstein, S. P. Harrison, and P. J. Bartlein. 2011. Modeling fire and
577   the terrestrial carbon balance. Global Biogeochemical Cycles **25**:Doi 10.1029/2010gb003906.
578 Pretzsch, H., R. Grote, B. Reineking, T. Rotzer, and S. Seifert. 2008. Models for forest ecosystem management: A
579   European perspective. Annals of Botany **101**:1065-1087.
580 Richardson, A. D., M. Williams, D. Y. Hollinger, D. J. P. Moore, D. B. Dail, E. A. Davidson, N. A. Scott, R. S. Evans, H.
581   Hughes, J. T. Lee, C. Rodrigues, and K. Savage. 2010. Estimating parameters of a forest ecosystem C model
582   with measurements of stocks and fluxes as joint constraints. Oecologia **164**:25-40.
583 Schmid, S., E. Thurig, E. Kaufmann, H. Lischke, and H. Bugmann. 2006. Effect of forest management on future
584   carbon pools and fluxes: A model comparison. Forest Ecology and Management **237**:65-82.
585 Schnute, J. T. 1991. The Importance of Noise in Fish Population-Models. Fisheries Research **11**:197-223.
586 Schroter, D., W. Cramer, R. Leemans, I. C. Prentice, M. B. Araujo, N. W. Arnell, A. Bondeau, H. Bugmann, T. R.
587   Carter, C. A. Gracia, A. C. de la Vega-Leinert, M. Erhard, F. Ewert, M. Glendining, J. I. House, S.
588   Kankaanpaa, R. J. T. Klein, S. Lavorel, M. Lindner, M. J. Metzger, J. Meyer, T. D. Mitchell, I. Reginster, M.
589   Rounsevell, S. Sabate, S. Sitch, B. Smith, J. Smith, P. Smith, M. T. Sykes, K. Thonicke, W. Thuiller, G. Tuck, S.
590   Zaehle, and B. Zierl. 2005. Ecosystem service supply and vulnerability to global change in Europe. Science
591   **310**:1333-1337.
592 Schwalm, C. R., C. A. Williams, K. Schaefer, R. Anderson, M. A. Arain, I. Baker, A. Barr, T. A. Black, G. S. Chen, J. M.
593   Chen, P. Ciais, K. J. Davis, A. Desai, M. Dietze, D. Dragoni, M. L. Fischer, L. B. Flanagan, R. Grant, L. H. Gu, D.
594   Hollinger, R. C. Izaurralde, C. Kucharik, P. Lafleur, B. E. Law, L. H. Li, Z. P. Li, S. G. Liu, E. Lokupitiya, Y. Q.
595   Luo, S. Y. Ma, H. Margolis, R. Matamala, H. McCaughey, R. K. Monson, W. C. Oechel, C. H. Peng, B. Poulter,
596   D. T. Price, D. M. Riciutto, W. Riley, A. K. Sahoo, M. Sprintsin, J. F. Sun, H. Q. Tian, C. Tonitto, H. Verbeeck,
597   and S. B. Verma. 2010. A model-data intercomparison of CO2 exchange across North America: Results
598   from the North American Carbon Program site synthesis. Journal of Geophysical Research-Biogeosciences
599   **115**:Doi 10.1029/2009jg001229.
600 Smith, P. J., G. D. Thornhill, S. L. Dance, A. S. Lawless, D. C. Mason, and N. K. Nichols. 2013. Data assimilation for
601   state and parameter estimation: application to morphodynamic modelling. Quarterly Journal of the Royal
602   Meteorological Society **139**:314-327.
603 Sullivan, A. L. 2009. Wildland surface fire spread modelling, 1990-2007. 3: Simulation and mathematical analogue
604   models. International Journal of Wildland Fire **18**:387-403.
605 Tildesley, M. J., R. Deardon, N. J. Savill, P. R. Bessell, S. P. Brooks, M. E. J. Woolhouse, B. T. Grenfell, and M. J.
606   Keeling. 2008. Accuracy of models for the 2001 foot-and-mouth epidemic. Proceedings of the Royal
607   Society B-Biological Sciences **275**:1459-1468.
608 Weng, E. S. and Y. Q. Luo. 2011. Relative information contributions of model vs. data to short- and long-term
609   forecasts of forest carbon dynamics. Ecological Applications **21**:1490-1505.
610 Weng, E. S. S., Y. Q. Luo, W. L. Wang, H. Wang, D. J. Hayes, A. D. McGuire, A. Hastings, and D. S. Schimel. 2012.
611   Ecosystem carbon storage capacity as affected by disturbance regimes: A general theoretical model.
612   Journal of Geophysical Research-Biogeosciences **117**:DOI: 10.1029/2012JG002040.
613 Williams, M., P. A. Schwarz, B. E. Law, J. Irvine, and M. R. Kurpius. 2005. An improved analysis of forest carbon
614   dynamics using data assimilation. Global Change Biology **11**:89-105.
615 Wu, X. W., Y. Q. Luo, E. S. Weng, L. White, Y. Ma, and X. H. Zhou. 2009. Conditional inversion to estimate
616   parameters from eddy-flux observations. Journal of Plant Ecology **2**:55-68.
617 Xia, J. Y., Y. Q. Luo, Y. P. Wang, and O. Hararuk. 2013. Traceable components of terrestrial carbon storage capacity
618   in biogeochemical models. Global Change Biology **19**:2104-2116.

619   Xu, T., L. White, D. F. Hui, and Y. Q. Luo. 2006. Probabilistic inversion of a terrestrial ecosystem model: Analysis of
620   uncertainty in parameter estimation and model prediction. Global Biogeochemical Cycles **20**:Doi
621   10.1029/2005gb002468.
622   Xue, H. D., F. Gu, and X. L. Hu. 2012. Data Assimilation Using Sequential Monte Carlo Methods in Wildfire Spread
623   Simulation. Acm Transactions on Modeling and Computer Simulation **22**:Doi 10.1145/2379810.2379816.
624   Zaehle, S., A. D. Friend, P. Friedlingstein, F. Dentener, P. Peylin, and M. Schulz. 2010. Carbon and nitrogen cycle
625   dynamics in the O-CN land surface model: 2. Role of the nitrogen cycle in the historical terrestrial carbon
626   balance. Global Biogeochemical Cycles **24**:Doi 10.1029/2009gb003522.
627   Zhou, T., P. J. Shi, D. F. Hui, and Y. Q. Luo. 2009. Global pattern of temperature sensitivity of soil heterotrophic
628   respiration (Q(10)) and its implications for carbon-climate feedback. Journal of Geophysical Research-
629   Biogeosciences **114**:Doi 10.1029/2008jg000850.
630   Zhou, X. H., T. Zhou, and Y. Q. Luo. 2012. Uncertainties in carbon residence time and NPP-driven carbon uptake in
631   terrestrial ecosystems of the conterminous USA: a Bayesian approach. Tellus Series B-Chemical and
632   Physical Meteorology **64**:Doi 10.3402/Tellusb.V3464i3400.17223.
633
634

635

Figure legends

Figure 1 Techniques framework for ecological prediction. Data assimilation uses observed data to train models and updates the model output so as to accurately describe the past and current states of the system and finally to better predict the future states.

Figure 2 Real-time forecasts of H1N1 in Singapore. The first three rows depict observations (crosses) and forecasts (grey shaded area) at three points in time during the outbreak (indicated by the triangle on the time axis). The cross in the lower left indicates an independent estimate of infection from adult seroconversion. The bottom panel shows the posterior absolute deviation between predicted and observed incidence. As data assimilation trained the model (going from the first row to the third row, the forecasts (gray shading) became more certain (row 3) and prediction error declined (row 4). Figure reproduced from Ong et al. 2010.
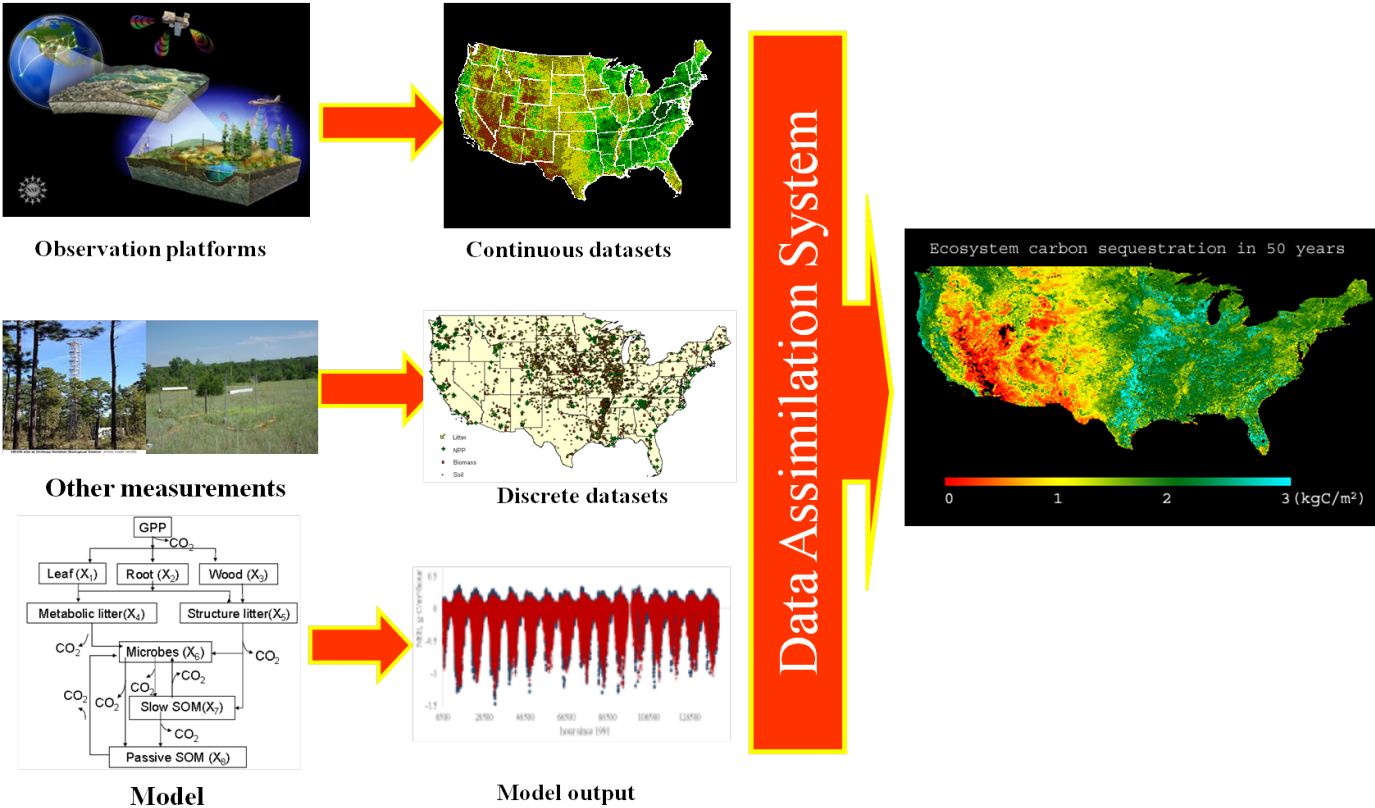
Figure 3 The variances of the estimated fishing mortalities for 7 years old fish using pure ensemble integration with no data assimilation (Ens pred) and three data assimilation techniques. Ensemble Kalman filter (EnKF), the ensemble smoother (ES), and ensemble Kalman smoother (EnKS) (Modified from Gronnevik & Evensen 2001). The figure shows that DA greatly reduces the variance of prediction.

Figure 4 Model uncertainty for the simulations of carbon fluxes (NEE, GPP, Ra, and Rh) for the FöBAAR model (Keenan et al. 2012). Three different approaches to constrain the model are shown: (1) using all data constrains (flux and biometric data) (black), (2) using short and long term flux data constrains (dark gray), and (3) using only short term flux data constrains (light

659   gray). The shaded areas represent the confidence in model predictions. The figure shows that the

660   model uncertainty can be quantified and reduced by assimilating different datasets into the model

661   (Modified from Keenan et al. 2012).

662

663   Figure 5 Daily analysis from1996-2004 (green) and daily forecast of carbon pools from 2004-

664   2012 (yellow) at Duke Forest with 100 ensembles after eight data sets were assimilated into the

665   TECO model using Ensemble Kalman Filter (EnKF). The uncertainties of the analysis were

666   reduced when data was assimilated into the model and the uncertainties of the forecasted carbon

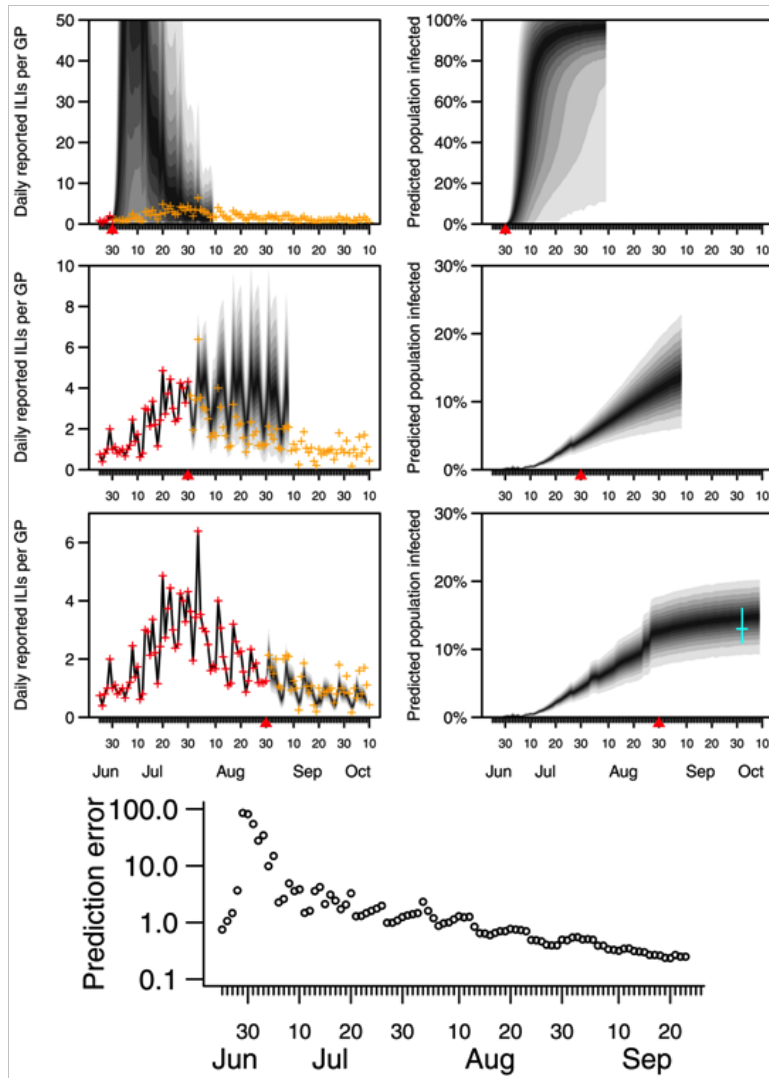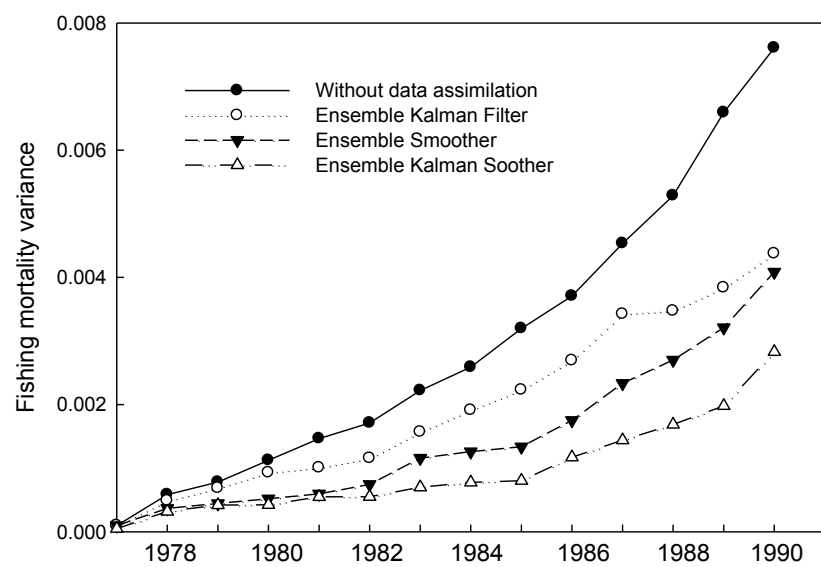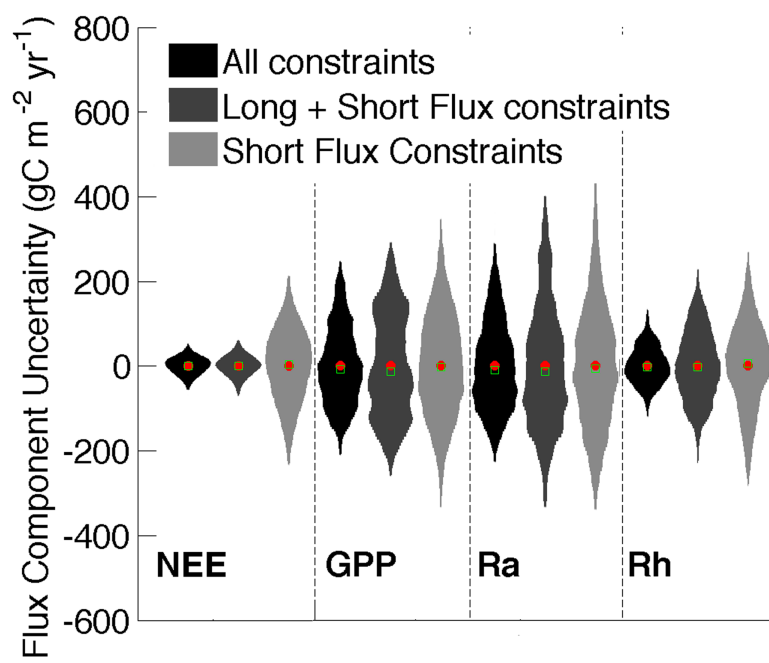667   pools were relatively stable (Modified from Gao et al. 2011).

668

669

Observation platforms

Continuous datasets

Other measurements

Discrete datasets

Model

Model output

Data Assimilation System

Ecosystem carbon sequestration in 50 years

0   1   2   3 (kgC/m²)

670

671   Figure 1

672

673

674 Figure 2

675
676   Figure 3

677

Figure 4

679

680   Figure 5

681