Ecological Informatics xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Data-driven techniques for modelling the gross primary production of the *páramo* vegetation using climate data: Application in the Ecuadorian Andean region

Veronica Minaya [a,b,c,]*, Gerald A. Corzo [a], Dimitri P. Solomatine [a,c], Arthur E. Mynett [a,c]

[a] *UNESCO-IHE, Institute for Water Education, Delft, The Netherlands*
[b] *Escuela Politécnica Nacional, Quito, Ecuador*
[c] *Technological University Delft, Delft, The Netherlands*

ARTICLE INFO

Article history:
Received 25 May 2015
Received in revised form 2 December 2016
Accepted 9 December 2016
Available online xxxx

Keywords:
Model tree
Neural network
k-nearest neighbour
Surrogate model
Vegetation models
BIOME-BGC

ABSTRACT

As one of the main areas of carbon cycle and climate change studies, water and CO_2 relations are of great significance for estimation of gross primary production (GPP). Various biogeochemical process-based models have been set up to estimate the GPP based on mathematical representation of biological, physiological and ecological processes. However, they ended up increasing the complexity and computational processing power due to the large number of physical equations that need to be solved. Computational time becomes an important matter in the simulation of multiple scenarios using models for long periods of time (e.g. climate projections). Data driven surrogate models have proven to be a useful tool for environmental modelling especially when ecological and climatic co-variates are large. The advantages of Data Driven Models (DDM) are: the possibility of adding new independent variables even if their understanding is weak, and short computational time to run. The aim is to explore the ability of DDMs to replicate a biochemical model calculating GPP. This study evaluates the performance of four surrogate DDMs, namely linear regression method (LRM), model tree (MT), instance-based learning (IBL) and artificial neural network (ANN). A simple empirical and semi-empirical relationship between GPP and climatic variables are studied. Input variable selection (IVS) methods were used to decide on the most relevant and potential environmental model inputs and then followed by a two-step approach which included a model-free and a model-based technique. Data from the highlands (*páramo* ecosystem) in the Ecuadorian Andean Region from 12-year time-series (2000 − 2011) were used to evaluate the models at various time frames and at different altitudes. The GPP time series data for the same period were derived from an earlier study using the biomodel BIOME-BGC (BioGeochemical Cycles), which is a comprehensive physical based model used in different analysis of carbon fluxes around the world. So-called IBL (nearest neighbour method) showed a great capability to reproduce the GPP when data was aggregated to monthly time frame. The computational time used to evaluate the time series with IBL as the selected DDM is shorter with enough accuracy for using it in multi-model runs.

1. Introduction

There is a growing interest in the estimation of terrestrial gross primary production (GPP) of ecosystems due to their role as sources or sinks of carbon and their contribution to the effects of climate change (Prentice et al., 2000). The GPP is the total amount of energy produced by the plants during photosynthesis and used for biomass production and respiration (Gough, 2011). GPP supports human well-being since it is the basis for food, fibre, wood production, and fuel. Additionally, GPP is one of the largest global CO_2 fluxes that controls several ecosystem functions (Beer et al., 2010); e.g. land-atmosphere interactions and carbon sequestration. Many process-oriented models have been proposed to deal with these complex interactions; however, in some cases, these models include a number of scientific hypotheses adopted for a particular ecosystem that might end up in an erroneous generalization when used for another ecosystem.

An attempt to estimate this difference at large grid cells was undertaken by Minaya et al. (2015a, 2016), showing an error of 23% on average if no spatial heterogeneity is considered. The mentioned studies have been carried out for *páramos*, a complex ‘hot spot’ mountain ecosystem that holds a great amount of biodiversity and unique ecological processes, thus providing important ecosystem services in terms of hydrological regulation and carbon storage.

Several vegetation and ecophysiological models have attempted to recreate the variation of GPP and evaluate the system behaviour (Cramer et al., 2001; McGuire et al., 2001). Validations have been carried out using carbon exchange monitoring measurements and also above and belowground biomass estimations, if available (Belgrano et al., 2001;

* Corresponding author at: UNESCO-IHE, Institute for Water Education, Delft, The Netherlands.
E-mail address: v.minayamaldonado@unesco-ihe.org (V. Minaya).

http://dx.doi.org/10.1016/j.ecoinf.2016.12.002
1574-9541/© 2016 Elsevier B.V. All rights reserved.

Hilbert and Ostendorf, 2001; Jung et al., 2007). However, there are some modelling issues that are difficult to resolve and which in most of the cases have been neglected leading to high uncertainties (Moorcroft, 2006; Morales et al., 2005). These refer to carbon monitoring in a consistent manner, approximations of nonexistent data, homogenization of plant functional types, static model parameters and site descriptors unchanged within an altitudinal gradient. On top of this, distributed process-oriented models can be computationally expensive: they typically have >30 parameters just to describe the vegetation processes.

In an attempt to reduce computational load during the model use, the use of surrogate models, i.e. simplified models (typically, data-driven) of process models, could be an alternative (Koziel and Leifsson, 2013; Regis and Shoemaker, 2013). However building them also requires building the process models first (with all the limitations mentioned above), and generation of large data sets for training the surrogate model requires multiple model runs leading to a serious computational effort as well. This is however done once, off-line, and multiple experiments with the resulting surrogate model do not require much time.

Data driven models (DDM) are constructed to represent complex interactions and allow data analysis, identification of trends and feasible predictions (Belgrano et al., 2001; Hilbert and Ostendorf, 2001; Papale and Valentini, 2003; Zhang et al., 2007). Basically a DDM is a (non-linear) statistical model describing the relationships between the input and output variables characterising the studied system. DMMs depend much less on theoretical assumptions, and in this regard are complementary to the process based models. DDM have limitations: they depend on the quality of the used data set and cannot generalise well for different conditions and use cases.

In ecological modelling numerous applications of a wide variety of DDM techniques have been reported, showing that it is possible to represent complex relationships which are not clearly explained by physically- or biologically-based considerations. In spatial dynamics, for instance, cellular automata and artificial neural networks have been applied for primary production (Anav et al., 2015; Belgrano et al., 2001; Scardi, 1996), carbon dioxide uptake and other carbon fluxes (Beer et al., 2010; Jung et al., 2011; Papale and Valentini, 2003; Xiao et al., 2014), radar forecasting (Li et al., 2013). DDMs have been also used for algae growth (Chen and Mynett, 2006; Li et al., 2010; Recknagel et al., 1997; Scardi, 1996), classification of landscape types (Brown et al., 1998; Zhang et al., 2007), distribution of vegetation (Hilbert and Ostendorf, 2001; Linderman et al., 2004) and hydrologic modelling for climate change scenarios (Corzo et al., 2009; Elshorbagy et al., 2010).

Looking specifically at terrestrial primary production, several studies have compared the use of data-driven methods such as multiple regression models and artificial neural networks (ANN) for a particular time frame (Jung et al., 2008; Papale and Valentini, 2003; Paruelo and Tomasel, 1997; Vetter et al., 2008). However, such examples are few and none of them have considered various time frames and a broader comparison of several DDMs. By comparing various time frames makes it possible to enhance the understanding of how influential the changes of temporal resolution and the selection of the precise meteorological variables are for building of the most adequate and accurate DDM.

This study evaluates the performance of data-driven model (DDM) techniques to discover the complex interactions between GPP and meteorological variables at various time frames. An existing BIOME-BGC model of the *páramo* Antisana was used as reference biomodel. Four DDMs where built as surrogate to simulate the GPP obtained from the biomodel, namely: linear regression method (LRM), model tree (MT), instance-based learning (IBL) and artificial neural network (ANN) model.

2. Material and methods

2.1. Case study

"Los Crespos - Humbolt" (LCH) basin, a small typical region in the Ecuadorian Andes, was selected to analyze the surrogate model capabilities to reproduce GPP. The LCH basin is located in the southwestern side of the volcano Antisana. It has an area of 15.2 km^2, of which 16% is covered by glacier, 17% by moraine and 68% with *páramo* vegetation and extends from 4010 m a.s.l. (meters above sea level) to 5000 m a.s.l. (Fig. 1). Precipitation range is 800–1200 mm/yr, monthly average temperature is 6 °C and average relative humidity is around 80%. Typical *páramo* vegetation covers the entire surface until the beginning of the moraine, which is mainly located at elevations above 4700 m a.s.l. In previous studies (Minaya et al., 2015a; Minaya et al., 2015b), plant species were identified and classified based on their growth forms (Ramsay and Oxley, 1997). In the lower and middle parts of the catchment the vegetation is dominated by tussock grasses (TU) (*Calamagrostis intermedia*) and acaulescent rosettes (AR) (*Werneria nubigena*, *Hypochaeris sessiliflora*). Near flood zones and streams there is a strong dominance of cushions (CU) (*Azorrella pedunculata*) (Minaya et al., 2015b). These growth forms had large differences in their carbon, nitrogen concentration and main ecophysiological characteristics along altitudinal gradients (Minaya et al., 2015b). For this reason, the parameters were adequately treated at three elevations (R1: 4000–4200 m a.s.l.; R2: 4200–4400 m a.s.l.; R3: 4400–4700 m a.s.l.).

2.2. Data description

Daily meteorological data were received from IRD (Institut de recherche pour le développement, Ecuador) and INAMHI (Instituto Nacional de Meteorología e Hidrología en Ecuador) databases. Daily total precipitation and daily maximum and minimum temperatures were collected from 2 stations, one in the upper and the other at the outlet of the basin for the years 2000–2011. The short wave radiation (SWR) and vapour pressure deficit (VPD) were calculated based on the above mentioned parameters by using a mountain climate simulator MT-CLIMB version 4.3 (Kimball et al., 1997; Running et al., 1987; Thornton, 2000; Thornton and Running, 1999) that estimate the near surface parameters based on nearby observations of temperature and precipitation. The use of VPD and SWR combined with the information of precipitation and temperature can be of certain value for the learning processes. All parameters were interpolated for the three different elevations (Table 1). There is no strong seasonality in the region, the spread of values of temperature in each altitudinal gradient are mainly attributed to the air humidity determined by local climate (Buytaert et al., 2006). Conversely to temperature, precipitation is highly variable in the region and at a small scale it is associated with wind speed and direction which in turn are controlled by slopes and irregular topography (Buytaert et al., 2005; Buytaert et al., 2006).

GPP is defined as the total amount of CO_2 that is fixed by the plants through photosynthesis and it has proved to be a good indicator of ecosystem's health, high GPP means high amount of CO_2 sequestration in the region and low values mean plant decay and organic matter decomposition. GPP was estimated using a biogeochemical and eco-physiological model BIOME-BGC (BioGeochemical Cycles), which is an ecosystem process model that estimates fluxes and storage of energy, water, carbon and nitrogen for soil and vegetation of terrestrial ecosystems (version 4.2; Thornton, 1998; Thornton et al., 2002). BIOME-BGC is a model capable of representing high complex ecological and biophysical process in ecosystems (White et al., 2000) and it has been successfully applied to estimate water and nutrient cycling from forest to herbaceous ecosystems (Di Vittorio et al., 2010; Hidy et al., 2012; Running and Hunt, 1993; Trusilova and Churkina, 2008). The model parameterization was done in a previous study (Minaya et al., 2015b) that relied on statistical analysis of key parameters derived from in situ measurements in order to reduce significantly the uncertainty of the calculated GPP. BIOME-BGC uses daily meteorological data and general stand soil information (Fig. 2) to simulate the energy, carbon, nitrogen and water cycles, and requires standard meteorological data as the main drivers for the ecosystem activity (Trusilova et al., 2009). The GPP for

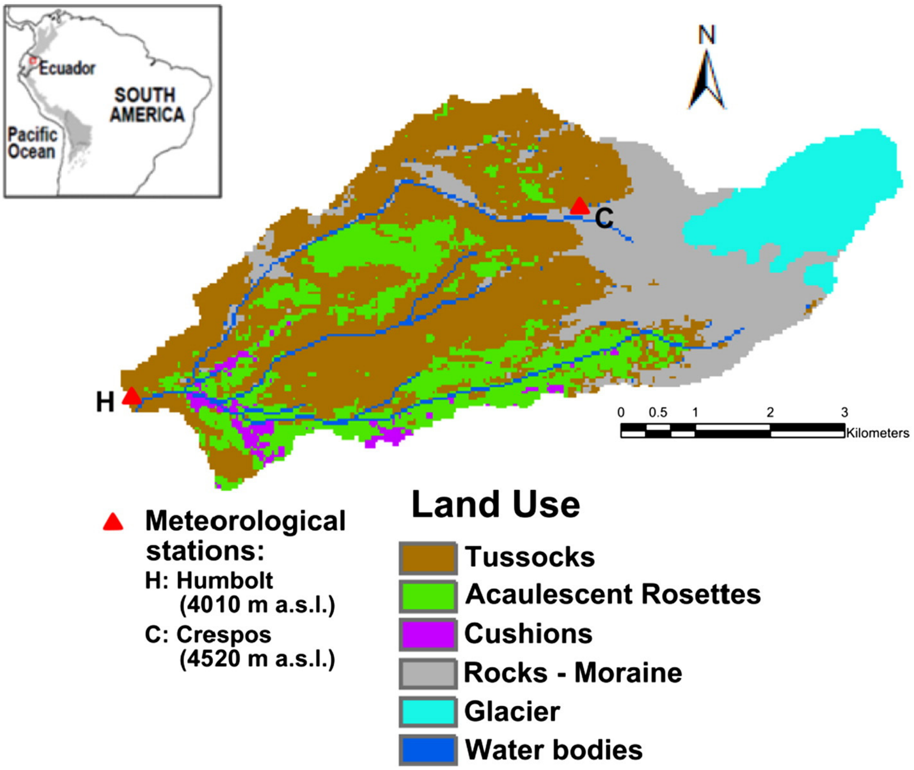

Fig. 1. Location and land use of the “Los Crespos – Humbolt” catchment, Ecuador.

each altitudinal range took into account the percentage of coverage of each growth form.

2.3. Data preparation

All climatic variables were aggregated at a monthly scale, and tested for simple monotonic trend using Mann-Kendall test (Gilbert, 1987; Kendall, 1975; Mann, 1945) over 12-year time series; there was no trend discovered ($p < 0.05$). Input variable selection (IVS) was used to decide on the most relevant and potential environmental model inputs (see e.g. Galelli et al. (2014)).

Data from three altitudinal ranges was used and grouped at various time frames (daily, weekly, bimonthly and monthly). The values of precipitation and GPP were aggregated while for other climatic variables an average function was used.

Table 1
Meteorological data summarize the climate file (annual means ± SD) based on 10 years of daily data (2000–2011). Where no reference is given, the value given was obtained from different sources (see text). Altitudinal ranges are R1: 4000–4200 m a.s.l.; R2: 4200–4400 m a.s.l.; R3: 4400–4700 m a.s.l.

Parameter	R1	R2	R3	Reference
Site and soil				
Elevation (m)	4100	4300	4500	–
Site latitude (°)	−0.4665			–
Albedo (DIM)	0.1723	0.1759	0.1753	–
Effective soil depth (m)	1.7	1.0	0.5	(Minaya et al., 2016)
Meteorological data				
Mean annual air temperature (°C)	7.31 ± 1.44	6.53 ± 0.35	4.82 ± 0.37	–
Mean annual precipitation (mm)	925.1 ± 100.8	1337.4 ± 196.0	1176.2 ± 184.7	–
Mean annual SWR (W/m^2)	0.47 ± 0.14	0.48 ± 0.14	0.45 ± 0.14	(Minaya et al., 2015a)
Mean annual VPD (Pa)	0.43 ± 0.15	0.39 ± 0.16	0.36 ± 0.13	(Minaya et al., 2015a)

2.4. Methodology

The methodology includes two main steps. At the first step, model-free analysis using criteria of correlation coefficient (CC) and average mutual information (AMI) between the climate time-series data and the GPP is carried out. At the second step, model-based analysis is used to test four different combinations of input variable sets, in which the lowest root mean square error (RMSE) leads to choose the best DDM. These four combinations are obtained following backward elimination defined by Blum and Langley (1997), where the search starts at a full selection of climatic variables and then variables are removed based on the performance and expert judgement. Bimonthly and monthly data sets were used to train LRM, MT, IBL and ANN model.

Normalized root mean squared error (NRMSE) was used as an indicator of the surrogate model performance. When looking at the bimonthly and monthly time frames, the lowest NRMSE values were shown in the monthly time frame.

2.5. DDMs set-up

For the development of surrogate DDMs, data was split in two data sets, 70% for training (from August 2003 to December 2011) and 30% for testing (from January 2000 to July 2003), following Elshorbagy et al. (2010). Table 2 shows that statistical properties of both data sets are similar.

The LRM and MT were built using the WEKA software (Witten and Frank, 2000). MT was built using the MP5 algorithm. It progressively splits the data set trying to ensure low standard deviation in subsets, and eventually generates several linear regression models for the resulting subsets (in the tree leaves). The minimum number of instances per leaf was set to 4.

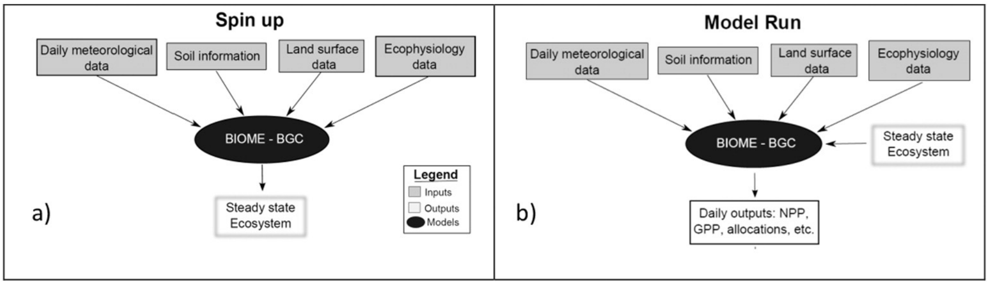

Fig. 2. Simplified scheme of the biomodeling strategy; a) Spin-up simulation; and b) Model run (12 years).

For instance-based learning (IBL) we used *k*-nearest neighbour algorithm for regression which was implemented in Matlab. The classification of the GPP was based on the Euclidean distance and calculated for each of the input vectors of the training set. The *k*-nearest neighbour algorithm weighted each of the *k* neighbours X_i based on their distance to the query point X_q, calculated as:

$$f(q) = \sum_{i=1}^{k} W_i f(X_i) / \sum_{i=1}^{k} W_i.$$

where W_i is a function of the distance between X_q and X_i. The weighing scheme tested was linear:

$$W_i = 1 - d(X_q, X_i)$$

The number of neighbours chosen was based on the minimum root mean squared error (RMSE) of the GPP from biomodel and the surrogate model. Since for large data sets IBL can be time-consuming, we limited our search to 30 nearest neighbours.

We employed an artificial neural network (ANN) with a multilayer perceptron (MLP) structure, with the logistic function in all nodes. Training was carried out by the back propagation algorithm. The number of nodes in the ANN model structure was determined by exhaustive model optimization, varying the number of nodes from 5 to 20, and selecting the best performance (root mean square error, RMSE) on training data. To prevent so-called network paralysis output data were normalized to a range [0,1]. The non-linear transfer logistic function was bind between 0.1 and 0.9 to avoid negative GPP values during the extrapolation. The ANN optimization algorithm used a regularized error (*msereg*) to reduce the overfitting and in general to improve the generalization of the neural network; it was calculated as follows:

$$msereg = \Upsilon * mse + (1 - \Upsilon) * msw$$

$$mse = \frac{1}{N} \sum_{i=1}^{N} (e_i)^2$$

$$msw = \frac{1}{n} \sum_{j=1}^{n} w_j^2,$$

where Y is the performance ratio, *mse* is the mean square error, *msw* is the mean square weight, e_i are the network errors and w_j are the network-weights.

We measured the performance of the techniques mentioned above: linear regression method (LRM), model tree (MT), instance-based learning (IBL) and artificial neural network (ANN). The model comparison was based on the normalized root mean square error (NRMSE) on the testing data set.

Taylor diagrams (Taylor, 2001) were used to compare the performance based on Pearson correlation, standard deviation and root mean square error (RMSE) of the surrogate models. These diagrams display pattern statistics; the radial distance from the centre of the graph is proportional to the standard deviation of a pattern. The Pearson correlation between the two models is given by the azimuthal position, to the right positive correlations and left negative correlations.

Fig. 3 shows the time-series data of GPP and meteorological values for monthly time frame at altitudinal range R2. These data were used in building all DDMs.

3. Results

3.1. Model-free IVS

The results of the application of the model-free approach for all altitudinal ranges are presented in Table 3, the input variables are all the climatic drivers (PRE, SWR, VPD, TEMP, TMAX, TMIN) that were considered important and linked to the variation of GPP. The highest values for the various time frames are shown in bold, bi-monthly and monthly time frames were used later for the analysis with different combinations of input variable sets as shown in Table 4.

Correlations were checked also with various time frames at the three elevations (Table 3), the highest correlations appear when data was aggregated in bimonthly and monthly time frames. For the low and mid elevations (R1 and R2, respectively), the GPP is

Table 2
Statistical properties of the training and testing data sets for all input and output variables.

	R1				R2				R3			
	Training		Testing		Training		Testing		Training		Testing	
Variable	μ	Σ	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
GPP	0.006	0.002	0.005	0.002	0.004	0.001	0.004	0.001	0.003	0.001	0.002	0.001
PRE	2.6	4.1	2.6	4.3	3.7	5.8	3.8	6.5	3.1	4.6	3.2	5.0
SWR	415.7	94.8	420.3	102.1	420.5	95.1	424.7	102.3	361.7	87.1	385.7	90.5
VPD	396.4	125.7	370.1	124.6	318.4	117.1	296.3	118.0	292.6	76.8	308.1	76.1
TEMP	7.6	1.6	7.2	1.5	6.4	1.5	6.2	1.5	4.8	1.4	4.8	1.3
TMAX	10.2	2.2	9.6	2.1	8.6	2.0	8.2	2.1	7.0	1.8	7.1	1.7
TMIN	0.8	1.7	0.7	1.8	0.7	1.8	0.9	1.8	−1.0	0.9	−1.4	0.8

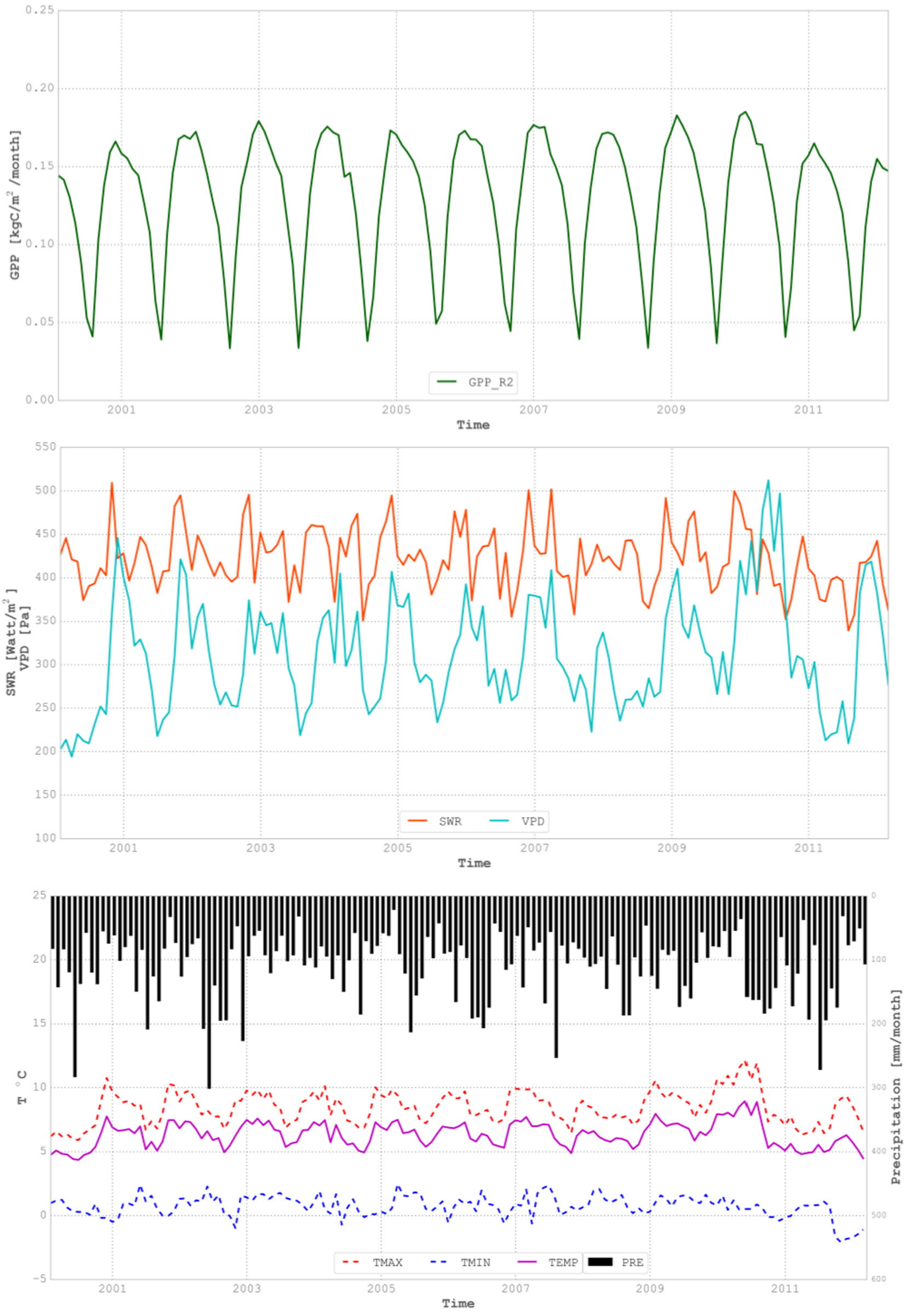

Fig. 3. Monthly time-series data of GPP and meteorological variables for 12 years (2000–2011) at altitudinal range R2 (4200–4400 m a.s.l.)

positively correlated to SWR (rho > 0.50, p < 0.05) to VPD (rho > 0.60, p < 0.01), mean and maximum temperature (rho > 0.55, p < 0.05 and rho > 0.60, p < 0.01 respectively). Conversely, precipitation and minimum temperature showed low values of correlation and AMI displaying a little direct relationship with the GPP. For higher elevations (R3) correlation was very low but it showed high AMI values, we assumed there was no good association between variables. Precipitation was further analysed but it did not show a straightforward relationship that contributes to the GPP variation. The input vector data that include SWR, VPD, TEMP and TMAX showed a strong influence on the GPP behaviour. Table 4 shows the input variables sets that were analysed.

Explicitly cross-validation in the LRM, MT and IBL was not performed; instead, generalization was ensured by the strategy of "early stopping of learning", i.e. deliberately undertraining the model and thus reducing accuracy on training set.

Table 3
Model-free IVS technique using criteria of correlation (CC) and average mutual information (AMI) between the climatic time-series data of precipitation (PRE), short wave radiation (SWR), vapour pressure deficit (VPD), mean temperature (TEMP), maximum temperature (TMAX), minimum temperature (TMIN) and gross primary production (GPP). Twelve years of time-series data (2000–2011) aggregated in four time frames: daily (1 d), weekly (7 d), bimonthly (15 d) and monthly (30 d) at three main altitudinal ranges: R1: 4000–4200 m a.s.l.; R2: 4200–4400 m a.s.l.; R3: 4400–4700 m a.s.l. Significant correlations ($p < 0.05$) and high values of AMI are indicated in bold.

			Model-free							
			CC				AMI			
Altitudinal ranges	Input variables	Output variable	1d	7d	15d	30d	1d	7d	15d	30d
R1	PRE	GPP	−0.16	−0.24	−0.30	−0.39	0.041	0.080	0.179	**0.336**
	SWR		0.34	0.40	**0.46**	**0.56**	0.144	0.146	0.232	**0.419**
	VPD		0.38	0.44	**0.46**	**0.49**	0.164	0.196	0.274	**0.479**
	TEMP		0.36	0.40	0.41	0.42	0.159	0.218	0.256	**0.413**
	TMAX		0.40	0.44	**0.45**	**0.47**	0.184	0.220	**0.308**	**0.436**
	TMIN		−0.11	−0.09	−0.10	−0.12	0.052	0.074	0.127	0.290
R2	PRE	GPP	−0.14	−0.21	−0.27	−0.34	0.017	0.075	0.154	**0.518**
	SWR		0.31	0.38	0.44	**0.53**	0.093	0.125	0.212	**0.561**
	VPD		0.38	0.44	**0.47**	**0.52**	0.127	0.210	0.258	**0.696**
	TEMP		0.40	**0.47**	**0.49**	**0.50**	0.158	0.238	**0.327**	**0.706**
	TMAX		0.42	**0.49**	**0.51**	**0.54**	0.168	0.250	**0.326**	**0.683**
	TMIN		−0.07	−0.04	−0.04	−0.05	0.029	0.067	0.126	**0.543**
R3	PRE	GPP	−0.11	−0.10	−0.10	−0.11	0.032	0.028	0.152	**0.483**
	SWR		0.31	0.32	0.34	0.38	0.110	0.073	0.227	**0.539**
	VPD		0.29	0.27	0.27	0.27	0.097	0.060	0.229	**0.585**
	TEMP		0.26	0.24	0.23	0.22	0.101	0.054	0.250	**0.637**
	TMAX		0.28	0.26	0.25	0.25	0.106	0.069	0.230	**0.600**
	TMIN		−0.04	−0.04	−0.04	−0.06	0.030	0.040	0.219	**0.556**

3.2. Comparison of DDMs

The input variable sets represent the combination of meteorological variables that affect the GPP variation. Notably, the input variables set 2 (which considers SWR, VPD, TEM and TMAX) stands out and show to be a good combination set of variables. Input variable set 3 slightly declines its performance emphasizing the sensitivity of SWR when it is not taken into consideration (Table 5). Major errors appear in the performance of ANN for both time frames. The final model structures are shown in Table 6.

We analysed the performance of each DDM using the input variables data set 2 as the one leading to the smallest NRMSE. The Taylor diagrams in Fig. 4 show the coefficient correlation, standard deviation and RMSE difference between the biomodel GPP and the surrogate data driven model. Performances of models for low and mid elevations are similar. IBL showed the capability to represent the GPP variation with minimum error, whereas ANN had the largest RMSE. For low and mid elevations (R1 & R2), IBL has the highest correlation coefficient for input variable set 2 ($r = 0.80$ and 0.84, respectively) and lowest RMSE difference ($RMSE = 0.037$ and 0.026, respectively). On the contrary, ANN does not perform well showing correlation coefficient $r < 0.6$ and a large RMSE difference ($RMSE > 0.045$). For higher elevations (R3), none of the DDM showed good results in terms of correlation coefficient and RMSE.

3.3. Order of effects

We compared the differences between the biomodel and surrogate model and assessed the advantage of using IBL using meteorological variables that change along an altitudinal gradient. The input variable set 2 was chosen as the best surrogate model to simulate GPP (smaller NRMSE). The final model structure is as follows:

$$GPP = f(SWR, \mathrm{VPD}, \mathrm{TEMP}, \mathrm{TMAX})$$

The IBL had better performance when data are aggregated in a monthly time frame. The prevailing pattern can be seen in Fig. 5, in which the surrogate model tends to underestimate the phenomena. It can harmonize the peaks and dips during the growth and decay season of the vegetation.

Table 4
The four input variables sets considered.

Set	Input variables	Output variable
1	PRE, SWR, VPD, TEMP, TMAX, TMIN	GPP
2	SWR, VPD, TEMP, TMAX	
3	VPD, TEM, TMAX	
4	VPD, TMAX	

3.4. Computational time

The biomodel BIOME-BGC used to estimate the GPP variation at each elevation took approximately 40 s for a 12-year time series. To run the biomodel for the entire basin, we might need >4 days to run pixels of 30 m × 30 m. During the DDM analysis, <1 s was needed for IBL to run training and testing for a 12-year time series data for all altitudinal

Table 5
NRMSE of DDM (LRM, MT, IBL, ANN) for GPP simulation at the three altitudinal ranges on the testing data set at a bimonthly (15 d), and monthly (30 d) time frame.

		R1		R2		R3	
Model	Input vars	15 d	30 d	15 d	30 d	15 d	30 d
LRM	Set 1	0.294	0.270	0.272	0.245	0.460	0.437
	Set 2	0.292	0.272	0.274	0.242	0.466	0.447
	Set 3	0.289	0.271	0.268	0.266	0.483	0.476
	Set 4	0.293	0.271	0.266	0.252	0.483	0.476
MT	Set 1	0.316	0.270	0.278	0.251	0.467	0.437
	Set 2	0.292	0.272	0.279	0.242	0.473	0.447
	Set 3	0.281	0.253	0.280	0.233	0.483	0.476
	Set 4	0.293	0.253	0.262	0.252	0.483	0.476
IBL	Set 1	0.282	0.259	0.276	0.222	0.475	0.464
	Set 2	0.280	0.237	0.276	0.208	0.482	0.462
	Set 3	0.274	0.238	0.269	0.213	0.501	0.493
	Set 4	0.274	0.238	0.269	0.214	0.501	0.493
ANN	Set 1	0.335	0.441	0.410	0.472	0.504	0.534
	Set 2	0.356	0.399	0.327	0.419	0.509	0.498
	Set 3	0.326	0.424	0.318	0.461	0.555	0.569
	Set 4	0.304	0.336	0.293	0.305	0.476	0.470

Table 6
Model structures: input variables sets, number of linear model for MT, neighbours for IBL and hidden nodes for ANN.

Model	Input	R1	R2	R3
MT	Set 1	3	3	1
	Set 2	3	3	1
	Set 3	3	1	1
	Set 4	1	3	2
IBL	Set 1	15	9	18
	Set 2	19	20	7
	Set 3	19	18	26
	Set 4	16	18	26
ANN	Set 1	5	6	5
	Set 2	5	11	5
	Set 3	5	5	6
	Set 4	5	6	5

ranges. This shows the usefullness of having DDM-based surrogate of a process-based model for experimentation and “what-if” analysis.

4. Discussion

4.1. GPP responses to climatic variables

The model-free technique used to identify and select the appropriate input variables showed the strongest relationships between GPP and each of the meteorological variables. However, it is difficult to attribute the variation of GPP to a single meteorological variable since most of them are indirectly dependent of each other (Jung et al., 2007). We found it crucial to investigate the strength and relationship between each of the climatic variables with GPP for uncertainty analysis in future climate projections

4.2. Surrogate model performance

IBL showed capability to represent the GPP fluctuations at a monthly time frame and it seemed to be able to capture the signal trend of the seasonal dynamics.

The biomodel BIOME-BGC is a continuous model that updates state variables while the DDMs look for the connections between the system variables without taking into account the explicit knowledge of the physical behaviour of the relationship between GPP and the climatic variables. There are some inconsistencies in the DDM output (e.g. larger error differences the beginning of the time series and the overestimation of GPP during the decay seasonal dynamics), but this could be explained by the particulars of the training data set rather than limitations of DDMs. Other studies (Levin, 1998; Scardi, 1996) have demonstrated that GPP fluctuations can be estimated just to a certain extent by conventional linear models due to the complexity of these terrestrial ecosystems. It would appear that it is equally important to analyze the scope at which the meteorological conditions perform on the GPP variation or whether the GPP is a consequence of the system's self-organization. In this regard, IBL can be used as quite accurate simulators of GPP variations besides the presence of a short and noisy time series data set.

The weak relationship between the climatic variables and the GPP discovered for higher elevations (R3) can be perhaps explained by the ambiguous responses and uncertainty in the model forcing fields when vegetation is located in the upper line close to the permanent snow. An in-depth analysis of the climatic data sets and the joint effects of them on the variations of GPP are beyond the scope of this study. However, an improved reanalysis of the meteorological forcing and its interpretation would reduce uncertainties in future long-term time-series data. We already considered the orographic factor that we believe it may have a stronger influence due to study area located in a mountainous region in the Ecuadorian Andes.

4.3. Performance based on variables selected and time frame

When looking at the outputs from the different sets of input data, the results highlighted the important role of SWR, VPD and temperature as climatic forcing processes on the representation of GPP. These variables are related to the vapour flow and transpirational demand which influences the amount of moisture that the plant tissues are exchanging with the atmosphere and therefore the capture of CO_2 that is entering the stomata.

The temporal resolution does play an important role in the changes of GPP especially in these high-altitudinal ecosystems, since they are characterized by slow decomposition and growth rates and therefore slow processes which agree with previous studies (Minaya et al., 2016; Spehn et al., 2006). This happens when the climatic variables are upscaled from daily to monthly values. It would appear that the GPP variation is not susceptible to sudden changes of high or low temperatures or presence of heavy rains at a regional scale.

4.4. Computational time

Current physically-based ecosystem models aim to create more realistic simulations but they ended up increasing the complexity and consuming lots of processing power. Additionally, determining appropriate values for the input data requires great diligence and for our case required an extensive fieldwork and analysis to get accurate values. Small uncertainties in the parameters may propagate to a wide range of variability in the simulations. As a follow up, simplifications are necessary but without compromising the capacity of the models. The surrogate model based on a data driven techniques was computationally faster and can easily be used to upscale and predict future climate scenarios from global climate data. In a simple exercise we compared the computational time used to estimate the GPP for one scenario of the Coupled Model Intercomparison Project (CMIP) models. In brief, to

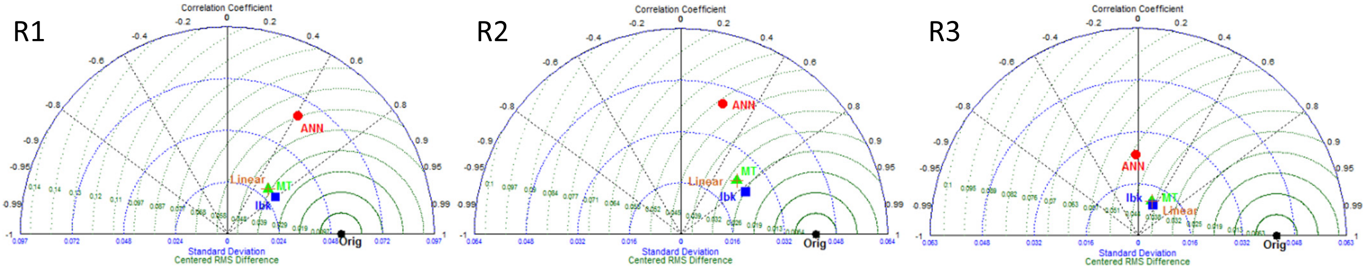

Fig. 4. Taylor diagram showing Pearson correlation, standard deviation and RMSE difference between GPP testing set and each of the DDM technique (LRM, MT, IBL, ANN) for a monthly time frame using input set 2 for all altitudinal ranges.

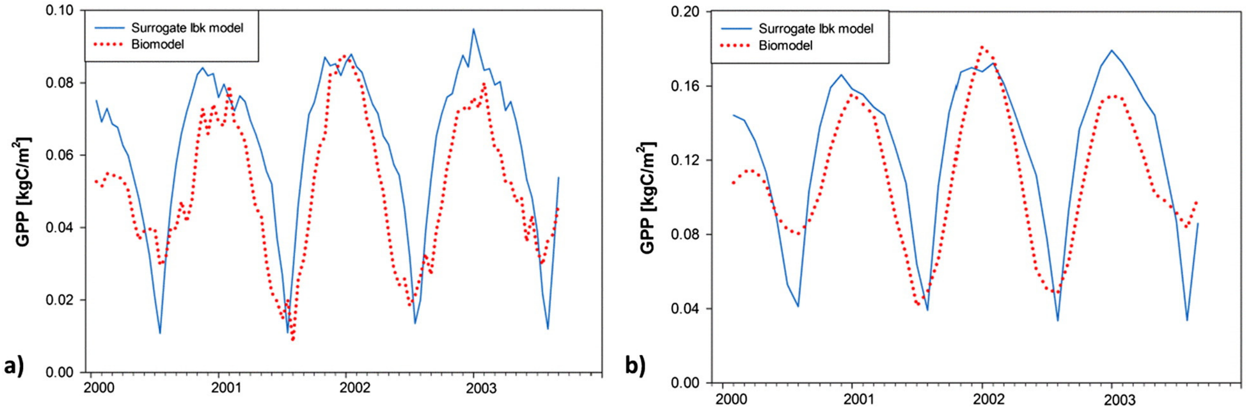

Fig. 5. Bio vs surrogate model for GPP simulation for mid-elevation R2 (4200–4400 m a.s.l.) using IBL for bimonthly and monthly time frames.

run a spatially distributed model of 10,300 cells using BIOME-BGC needs approximately 320 days in comparison to the almost 100 times faster IBL. The ecosystem process model requires a substantial investment of computational time in contrast to the DDM, which is shorter with enough accuracy for using it in multi-model runs.

5. Conclusions

This paper explored the use of surrogate DDMs to simulate the GPP along an altitudinal gradient in the *páramos* ecosystem in the Ecuadorian Andes. This was done by identifying the best input variables set for different time frames, for which the two methods were used - model free and model based. Based on the selected inputs, several models have been built (LM, MT, IBL, ANN), and for the considered use case, IBL model showed the best ability overall to reproduce the biomodel across a continuum of temporal scales. It was more responsive and sensitive to SWR, VPD, TEMP and TMAX as the main drivers for GPP on a monthly basis. However other models have also showed reasonable performance and all of them can be recommended for use in similar circumstances.

The IBL surrogate model does not replace a detailed and comprehensive physical based model, but it is a complementary statistical technique that link input and output variables for a comprehensive analysis of the ecosystem. The short computational time in running of the IBL will allow the extrapolation to higher temporal scales in future estimations of gross primary production (GPP), especially when using climate change scenarios such as the CMIP (Coupled Model Intercomparison Project). Fast running surrogate models allow for experimentation and "what-if" analysis. In case of availability of more observations DDMs can be built not on the basis of data generated by the process model, but directly on measurements, and this paper confirms that it is possible.

Although the DDM techniques tested in this paper showed that precipitation was not a variable that influence the variation of GPP, it is well known that precipitation is the major driving force for plant growth and therefore carbon uptake by plants. The ability of the DDM techniques to model the climate scenarios and the sensitivity of GPP to precipitation deserves further studies due to the high number of complex biological processes (i.e. adaptation to climate, to nutrients availability and others).

Further research will focus on a more detailed analysis of how frequency, timing and amount of precipitation would influence thresholds for carbon uptake by the grassland and therefore GPP in the region, rather than using the total precipitation as investigated in this paper. Parameters such as leaf area index and available nitrogen content in soil are equally important and could be included within the input set of parameters. Since the photosynthesis is the main process for primary production which is also driven by indirect controls that operate environmental conditions on the mineralization of nitrogen due to litter and soil organic matter decomposition (Running et al., 2000).

Acknowledgement

Financial support came from SENESCYT (Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación) and from the Dutch Ministry of Foreign Affairs (DUPC program at UNESCO-IHE). Other institutions that also cooperated in the provision of key information and data are EPMAPS, EPN, INAMHI, INIGEMM, Ministry of Environment, Herbario Nacional del Ecuador. The opinions expressed herein are those of the author(s) and do not necessarily reflect views of any of the Institutions named above.

References

Anav, A., Friedlingstein, P., Beer, C., Ciais, P., Harper, A., Jones, C., Murray-Tortarolo, G., Papale, D., Parazoo, N.C., Peylin, P., Piao, S., Sitch, S., Viovy, N., Wiltshire, A., Zhao, M., 2015. Spatiotemporal patterns of terrestrial gross primary production: a review. Rev. Geophys. 53, 785–818.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, A., Baldocchi, D., Bonan, G.B., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, K.W., Roupsard, O., Veenendaal, E., Viovy, N., Williams, C., Woodward, I., Papale, D., 2010. Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. Science 329, 834–838.

Belgrano, A., Malmgren, B.A., Lindahl, O., 2001. Application of artificial neural networks (ANN) to primary production time-series data. J. Plankton Res. 23, 651–658.

Blum, A.L., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artif. Intell. 97, 245–271.

Brown, D.G., Lusch, D.P., Duda, K.A., 1998. Supervised classification of types of glaciated landscapes using digital elevation data. Geomorphology 21, 233–250.

Buytaert, W., Iniguez, V., Celleri, R., De Bievre, B., Wyseure, G., Deckers, J., 2005. Analysis of the water balance of small paramo catchments in south Ecuador. International Conference on Headwater Control VI: Hydrology, Ecology and Water Resources in Headwaters (ed by Bergen, Norway, 20–23 June).

Buytaert, W., Celleri, R., De Bievre, B., Hofstede, R., Cisneros, F., Wyseure, G., Deckers, J., 2006. Human impact on the hydrology of the Andean paramos. Earth-Sci. Rev. 79, 53–72.

Chen, Q., Mynett, A.E., 2006. Modelling algal blooms in the Dutch coastal waters by integrated numerical and fuzzy cellular automata approaches. Ecol. Model. 199, 73–81.

Corzo, G.A., Solomatine, D.P., Hidayat, de Wit, M., Werner, M., Uhlenbrook, S., Price, R.K., 2009. Combining semi-distributed process-based and data-driven models in flow simulation: a case study of the Meuse river basin. Hydrol. Earth Syst. Sci. 13, 1619–1634.

Cramer, W., Bondeau, A., Woodward, F., Prentice, I.C., Betts, R.A., Brovkin, V., Cox, P.M., Fisher, V., Foley, J.A., Friend, A.D., Kucharik, C., Lomas, M.R., Ramankutty, N., Sitch, S., Smith, B., White, A., Young-Molling, A., 2001. Global response of terrestrial ecosystem structure and function to CO_2 and climate change: results from six dynamic global vegetation models. Glob. Chang. Biol. 7, 357–373.

Di Vittorio, A.V., Anderson, R.S., White, J.D., Miller, N.L., Running, S.W., 2010. Development and optimization of an Agro-BGC ecosystem model for C4 perennial grasses. Ecol. Model. 221, 2038–2053.

Elshorbagy, A., Corzo, G., Srinivasulu, S., Solomatine, D.P., 2010. Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology - part 1: concepts and methodology. Hydrol. Earth Syst. Sci. 14, 1931–1941.

Galelli, S., Humphrey, G., Maier, H., Castelletti, A., Dandy, G., Gibbs, M., 2014. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ. Model. Softw. 62, 33–51.

Gilbert, R.O., 1987. Statistical Methods for Environmental Pollution Monitoring. Wiley, New York.

Gough, C.M., 2011. Terrestrial primary production: fuel for life. Nat. Educ. 3, 28.

Hidy, D., Barcza, Z., Haszpra, L., Churkina, G., Pintér, K., Nagye, Z., 2012. Development of the biome-BGC model for simulation of managed herbaceous ecosystems. Ecol. Model. 226, 99–119.

Hilbert, D.W., Ostendorf, B., 2001. The utility of artificial neural networks for modelling the distribution of vegetation in past, present and future climates. Ecol. Model. 146, 311–327.

Jung, M., Verstraete, M., Gobron, N., Reichstein, M., Papale, D., Bondeau, A., Robustelli, M., Pinty, B., 2008. Diagnostic assessment of European gross primary production. Glob. Chang. Biol. 14, 2349–2364.

Jung, M., Vetter, M., Herold, M., Churkina, G., Reichstein, M., Zaehle, S., Ciais, P., Viovy, N., Bondeau, A., Chen, Y., Trusilova, K., Feser, F., Heimann, M., 2007. Uncertainties of modeling gross primary productivity over Europe: a systematic study on the effects of using different drivers and terrestrial biosphere models. Glob. Biogeochem. Cycles 21.

Jung, M., Reichstein, M., Margolis, H.A., et al., 2011. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. J. Geophys. Res.:116 http://dx.doi.org/10.1029/2010jg001566.

Kendall, M.G., 1975. Rank Correlation Methods. fourth ed. Charles Griffin, London.

Kimball, J.S., Running, S.W., Nemani, R., 1997. An improved method for estimating surface humidity from daily minimum temperature. Agric. For. Meteorol. 85, 87–98.

Koziel, S., Leifsson, L. (Eds.), 2013. Suroagte-based Modeling and Optimization: Applications in Engineering. Springer.

Levin, S.A., 1998. Ecosystems and the biosphere as complex adaptive systems. Ecosystems 1, 431–436.

Li, H., Corzo Perez, G., Martinez, C., Mynett, A.E., 2013. Self-learning cellular automata for forecasting precipitation from radar images. Hydrol. Eng. 18, 206–211.

Li, H., Arias, M., Blauw, A., Los, H., Mynett, A.E., Peters, S., 2010. Enhancing generic ecological model for short-term prediction of southern North Sea algal dynamics with remote sensing images. Ecol. Model. 221, 2435–2446.

Linderman, M., Liu, J., Qi, J., An, L., Ouyang, Z., Yang, J., Tan, Y., 2004. Using artificial neural networks to map the spatial distribution of understorey bamboo from remote sensing data. Int. J. Remote Sens. 25, 1685–1700.

Mann, H.B., 1945. Non-parametric tests against trend. Econometrica 13, 163–171.

McGuire, A.D., Sitch, S., Clein, J.S., Dargaville, R., Esser, G., Foley, J., Heimann, M., Joos, F., Kaplan, J., Kicklighter, D.W., Meier, R.A., Melillo, J.M., Moore, B., Prentice, C., Ramankutty, N., Reichenau, T., Schloss, A., Tian, H., Williams, L.J., Wittenberg, U., 2001. Carbon balance of the terrestrial biosphere in the twentieth century: analyses of CO_2, climate and land use effects with four process-based ecosystem models. Glob. Biogeochem. Cycles 15, 183–206.

Minaya, V., Corzo, G., van der Kwast, J., Mynett, A.E., 2016. Simulating gross primary production and stand hydrological processes of páramo grasslands in the Ecuadorian Andean Region using BIOME-BGC model. Soil Sci. 181 (7):335–346. http://dx.doi.org/10.1097/SS.0000000000000154.

Minaya, V., Corzo, G., Van der Kwast, J., Galarraga-Sanchez, R., Mynett, A.E., 2015a. Classification and multivariate analysis of differences in gross primary production at different elevations using BIOME-BGC in the páramos; Ecuadorian Andean Region. Revista de Matemática: Teoría y aplicaciones vol. 22, No. 2. CIMPA, San Jose - Costa Rica, pp. 369–394 ISSN: 1409-2433.

Minaya, V., Corzo, G., Romero-Saltos, H., van der Kwast, J., Lantinga, E., Galarraga-Sanchez, R., Mynett, A.E., 2015b. Altitudinal analysis of carbon stocks in the Antisana *páramo*, Ecuadorian Andes. Plant Ecol.:1–11 http://dx.doi.org/10.1093/jpe/rtv073.

Moorcroft, P.R., 2006. How close are we to a predictive science of the biosphere? Trends Ecol. Evol. 21, 400–407.

Morales, P., Sykes, M.T., Prentice, I.C., Smith, P., Smith, B., Bugmann, H., Zierl, B., Friedlingstein, P., Viovy, N., Sabate, S., Sanchez, A., Pla, E., Gracia, C.A., Sitch, S., Arneth, A., Ogee, J., 2005. Comparing and evaluating process-based ecosystem model predictions of carbon and water fluxes in major European forest biomes. Glob. Chang. Biol. 11, 2211–2233.

Papale, D., Valentini, R., 2003. A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. Glob. Chang. Biol. 9, 525–535.

Paruelo, J.M., Tomasel, F., 1997. Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models. Ecol. Model. 98, 173–186.

Prentice, I.C., Heimann, M., Sitch, S., 2000. The carbon balance of the terrestrial biosphere: ecosystem models and atmospheric observations. Ecol. Appl. 10, 1553–1573.

Ramsay, P.M., Oxley, E.R.B., 1997. The growth form composition of plant communities in the Ecuadorian páramos. Plant Ecol. 131, 173–192.

Recknagel, F., French, M., Harkonen, P., Yabunaka, K., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecol. Model. 96, 11–28.

Regis, R., Shoemaker, C., 2013. Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization. Eng. Optim. 45, 529–555.

Running, S.W., Hunt, E.R. (Eds.), 1993. Generalization of a forest ecosystem process model for other biomes, BIOME-BGC, and an application for global-scale models. In: "Scaling physiological processes: leaf to globe", pp. 141–158 (San Diego, CA, USA).

Running, S.W., Nemani, R.R., Hungerford, R.D., 1987. Extrapolation of synoptic meteorological data in mountainous terrain and its use for simulating forest evaporation and photosynthesis. J. Forest. Res. 17, 472–483.

Running, S.W., Thornton, P.E., Nemani, R., Glassy, J.M. (Eds.), 2000. Global Terrestrial Gross and Net Primary Productivity from the Earth Observing System. Springer-Verlag New York.

Scardi, M., 1996. Artificial neural networks as empirical models for estimating phytoplankton production. Mar. Ecol. Prog. Ser. 139, 289–299.

Spehn, E.M., Liberman, M., Korner, C. (Eds.), 2006. Land Use Change and Mountain Biodiversity.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. 106, 7183–7192.

Thornton, P.E., 1998. Regional Ecosystem simulation: combining surface and satellite based observations to study linkages between terrestrial energy and mass budgets. PhD thesis. School of Forestry, University of Montana, Missoula (280 pp.).

Thornton, P.E., 2000. Simultaneous estimation of daily solar radiation and humidity from observed temperature and precipitation: an application over complex terrain in Austria. Agric. For. Meteorol. 104, 255–271.

Thornton, P.E., Running, S.W., 1999. An improved algorithm for estimating incident daily solar radiation from measurements of temperature, humidity, and precipitation. Agric. For. Meteorol. 93, 211–228.

Thornton, P.E., Law, B.E., Gholz, H.L., Clark, K.L., Falge, E., Ellsworth, D.S., Goldstein, A.H., Monson, R.K., Hollinger, D., Flak, M., Chen, J., Sparks, J.P., 2002. Modeling and measuring the effects of disturbance history and climate on carbon and water budgets in evergreen needleleaf forests. Agric. For. Meteorol. 113, 185–222.

Trusilova, K., Churkina, G., 2008. The Terrestrial Ecosystem Model GBIOME-BGCv1. Max-Planck Institute for Biogeochemistry, Jena, Germany.

Trusilova, K., Trembath, J., Churkina, G., 2009. Parameter estimation and validation of the terrestrial ecosystem model BIOME-BGC using eddy-covariance flux measurements. T.R. 16. Max Planck Institut fur Biogeochemie, Jena.

Vetter, M., Churkina, G., Jung, M., Reichstein, M., Zaehle, S., Bondeau, A., Chen, Y., Ciais, P., Feser, F., Freibauer, A., Geyer, R., Jones, C., Papale, D., Tenhunen, J., Tomelleri, E., Trusilova, K., Viovy, N., Heimann, M., 2008. Analyzing the causes and spatial pattern of the European 2003 carbon flux anomaly using seven models. Biogeosciences 5, 561–583.

White, M.A., Thornton, P.E., Running, S.W., 2000. Parameterization and sensitivity analysis of the BIOME-BGC terrestrial ecosystem model: net primary production controls. Earth Interact. 4, 1–85.

Witten, I.H., Frank, E., 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmannp.

Xiao, J., Ollinger, S.V., Frolking, S., Hurtt, G.C., Hollinger, D.Y., Davis, K.J., Pane, Y., Zhang, X., Deng, F., Chen, J., Baldocchi, D.D., Law, B.E., Arain, M.A., Desai, A.R., Richardson, A.D., Sunn, G., Amiro, B., Margolis, H., Gu, L., Scott, R.L., Blanken, P.D., Suyker, A.E., 2014. Data-driven diagnostics of terrestrial carbon dynamics over North America. Agric. For. Meteorol. 197, 142–157.

Zhang, Z., Verbeke, L., Clercq, E., Ou, X., Wulf, R., 2007. Vegetation change detection using artificial neural networks with ancillary data in Xishuangbanna, Yunnan Province, China. Chin. Sci. Bull. 52, 232–243.