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/

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

Please cite this article as: Minaya, V., et al., Data-driven techniques for modelling the gross primary production of the *páramo* vegetation using climate data: Application in the Ecuadorian Andean region, Ecological Informatics (2016), http://dx.doi.org/10.1016/j.ecoinf.2016.12.002