Environmental Modelling & Software 41 (2013) 231–238

Contents lists available at SciVerse ScienceDirect

Environmental Modelling & Software

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

Large-scale, high-resolution agricultural systems modeling using a hybrid approach combining grid computing and parallel processing

Gang Zhao [a,b,*], Brett A. Bryan [b], Darran King [b], Zhongkui Luo [c], Enli Wang [c], Ulrike Bende-Michl [c], Xiaodong Song [b,e], Qiang Yu [a,d]

[a] Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, 11A Datun Road, Anwai, Beijing 100101, China
[b] CSIRO Ecosystem Sciences, Waite Campus, Urrbrae, SA 5064, Australia
[c] CSIRO Land and Water, Black Mountain, Canberra, ACT 2601, Australia
[d] Plant Functional Biology and Climate Change Cluster, School of the Environment, University of Technology, Sydney, Broadway 2007 NSW, PO Box 123, Australia
[e] Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China

A R T I C L E I N F O

Article history:
Received 17 January 2012
Received in revised form
7 August 2012
Accepted 20 August 2012
Available online 10 September 2012

Keywords:
Grid computing
Parallel programming
APSIM
Agricultural modeling
High-performance computing
Food security
Crop modeling

A B S T R A C T

The solution of complex global challenges in the land system, such as food and energy security, requires information on the management of agricultural systems at a high spatial and temporal resolution over continental or global extents. However, computing capacity remains a barrier to large-scale, high-resolution agricultural modeling. To model wheat production, soil carbon, and nitrogen dynamics in Australia's cropping regions at a high resolution, we developed a hybrid computing approach combining parallel processing and grid computing. The hybrid approach distributes tasks across a heterogeneous grid computing pool and fully utilizes all the resources of computers within the pool. We simulated 325 management scenarios (nitrogen application rates and stubble management) at a daily time step over 122 years, for 12,707 climate–soil zones using the Windows-based Agricultural Production Systems SIMulator (APSIM). These simulations would have taken over 30 years on a single computer. Our hybrid high performance computing (HPC) approach completed the modeling within 10.5 days—a speed-up of over 1000 times—with most jobs finishing within the first few days. The approach utilizes existing idle organization-wide computing resources and eliminates the need to translate Windows-based models to other operating systems for implementation on computing clusters. There are however, numerous computing challenges that need to be addressed for the effective use of these techniques and there remain several potential areas for further performance improvement. The results demonstrate the effectiveness of the approach in making high-resolution modeling of agricultural systems possible over continental and global scales.

Software availability

Software: Grid-Parallel-APSIM
Developer: Gang Zhao
Contact address: Gang Zhao, CSIRO Sustainable Ecosystems PMB 2, Waite Campus, Urrbrae, SA 5064, Australia. Gang.Zhao@csiro.au
Software and hardware requirements: APSIM, Windows XP or higher, Condor, python, multi-core computer or CPU cluster
Language: Python 2.7
Availability: The software is free to use for educational and research purposes.

1. Introduction

Agricultural land plays a key part in the global issues of food and energy security (Foley et al., 2011). Pressure on land resources is expected to continue growing in the coming decades as a result of expanding population, changing food consumption patterns, and competition from alternative land uses such as biofuel feedstocks (Foley et al., 2011; Tilman et al., 2009). Risk factors such as climate change will continue to challenge agricultural productivity in the major agricultural regions of the world (Luo et al., 2005a). The response of agricultural systems to these drivers needs to be

* Corresponding author. CSIRO Ecosystem Sciences PMB 2, Waite Campus, Urrbrae, SA 5064, Australia. Tel.: +61 08 8303 8679, fax: +61 08 8303 8582.
E-mail address: Gang.Zhao@csiro.au (G. Zhao).

1364-8152/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.envsoft.2012.08.007

understood and predicted to inform policy for managing land resources and increasing the resilience of the land system to various risk factors (Bryan et al., 2010).

Process-based models such as the Agricultural Production Systems SIMulator (APSIM) (Keating et al., 2003) and Environmental Policy Integrated Climate Model (EPIC) (Liu, 2009; Williams et al., 1989) have been increasingly used to simulate aspects of agricultural systems including yields, soil organic carbon, water use efficiency, nitrogen use efficiency, greenhouse gas emissions, and energy balance (Gaiser et al., 2010; Luo et al., 2011; Paterson and Bryan, in press). Responses of agricultural systems to changes in external drivers such as management and climate have also been predicted (Luo et al., 2005a, 2007). Whilst most of these models have been designed for and used in simulating plant-environment processes at high temporal resolution (e.g. daily time step) at the plot scale, this information is required over large extents to inform policy. Drivers of the agricultural processes of yield and soil carbon—such as soil and climatic conditions—vary across the landscape (Hansen and Jones, 2000; Luo et al., in review). Equally, agricultural management practices need to be assessed for their impact on agricultural systems at a fine level of granularity (i.e. what fertilizer/pesticides to apply, how much and when to apply them, what cultivars to use etc.). The influence of these practices also varies with soil and climatic conditions (Akponikpè et al., 2010; Basso et al., 2010; Goulding et al., 2008; Zhao et al., in review). Thus, accurate representations of agricultural systems for addressing the challenges discussed above require the exploration of a high-dimensional management scenario space (Smit and Skinner, 2002), at high spatial resolution (Bryan et al., 2011; Folberth et al., 2012), over large areas (Wang et al., 2009). However, applying high-resolution spatio-temporal process-based models over large extents presents significant computational challenges (Nichols et al., 2011). Whilst a single plot-based scenario may be completed within just minutes, many hundreds of scenarios may be required over many thousands of spatial units. This cannot be done within an acceptable time period using traditional computing methods.

One way to meet this computational demand is to process the simulations in parallel using high-performance computing approaches (Wang et al., 2011). Cluster and grid computing are the most common approaches. Clusters use a collection of linked, homogenous computers working together as a single system with tasks scheduled through job management software. Nichols et al. (2011) demonstrated the potential of computing clusters for modeling agricultural systems, achieving a 40 times speedup in running 140,000 EPIC simulations concurrently on a Linux-based computing cluster. However, most agricultural models are built for the Windows operating systems rather than Linux which is the most common operating system for large clusters, and non-trivial costs are associated with translating the software to another operating system. Grid computing can offer a viable alternative to clusters for high-throughput computing without the need to translate Windows-based models to another operating system.

Large organizations can have many Windows-based desktop computers connected through high-speed networks which, with significant idle time, commonly operate at only a fraction of their processing potential (Huang and Yang, 2011). A key advantage of grid computing is that it can effectively coordinate loosely coupled, heterogeneous, and geographically dispersed computing resources over multiple administrative domains to achieve a common computing goal (Jeffery, 2007; Schwiegelshohn et al., 2010). Grid computing is highly anticipated by those facing compute-intensive problems with no access to local clusters (Sulis, 2009) and has been demonstrated to achieve significant computation efficiencies in environmental modeling (Fernández-Quiruelas et al., 2011; Mineter et al., 2003; Sulis, 2009).

Grid computing attains significant computing efficiencies through coordinating many idle computers. However, grid computing alone can still fail to utilize the full processing capacity of individual nodes in the grid pool. Although many modern computer workstations are equipped with multi-core CPUs, most grid middleware can only allocate one serial application to each node (Huang and Yang, 2011). Parallel programming methods can run multiple instructions simultaneously and take advantage of multi-core hardware and accelerate processing in proportion to the number of CPU cores (Hillar, 2010; Rouholahnejad et al., 2012). Parallel programming has been frequently demonstrated to improve the computing efficiency of models (Elaine, 2005; von Bloh et al., 2010). Embedding parallel programming techniques into grid computing could also substantially improve the throughput of grid computing.

In this study, we confronted a computing challenge of executing more than four million simulations of the APSIM process-based agricultural systems model. Processing each simulation takes 4 min on a single CPU core. We simulated wheat productivity on a daily time step over 122 years for 325 management scenarios in 12,707 spatial units covering Australia's cereal cropping regions. A hybrid computing approach was developed to distribute tasks to idle computers in a computing grid via the middleware, Condor. The approach also employed parallel processing methods to improve the throughput of grid computing. The agricultural systems modeling context and workflow of the hybrid computing approach is provided in the Methods section. The performance of the hybrid computing approach is presented in the Results section in addition to some illustrative model outputs. We demonstrate the potential for high spatial and temporal resolution modeling of agricultural systems over very large spatial extents for addressing global challenges affecting the land system such as food and energy security.

2. Methods

2.1. Study area and spatial units

A range of biophysical factors related to soil and climate determine the suitability of land for growing wheat (Nicholls, 1997). Suitable conditions prevail on mainland Australia in an area west of the Great Dividing Range from Central Queensland through New South Wales and Victoria and on to South Australia and in the south west of Western Australia. The study area was the area currently under wheat production buffered by 100 km (ABARE, 2006). We divided the study area into climate–soil units (CS units) that are relatively homogenous at the scale of analysis. To define unique climate–soil combinations (CS units), we identified climate domains by *k*-means clustering (Hartigan and Wong, 1979) of relevant climate factors (rainfall; minimum, average and maximum temperature) and overlaid them with a soil type classification layer (Johnston et al., 2003) (Fig. 1). Spatial resolution is a key decision in agricultural systems modeling over large extents (Liu et al., 2007; Luo et al., 2005b; Zwart et al., 2010). Selecting a spatial resolution for modeling depends on the questions asked and the scale and confidence in available input data (Adam et al., 2011; Folberth et al., 2012). Too low a resolution causes homogenization of areas with meaningful spatial variation. Too high a resolution could lead to over fitting, a lack of confidence in input data, and present increased computational demand for little gain in modeling detail. The climate–soil domain separation method presented here used irregular polygons to define basic modeling units thereby eliminating the potential redundancies of using raster tessellation (Liu, 2009; Reidsma et al., 2009). This captured the spatial resolution contained in soil and climate data whilst minimizing the number of spatial units and resultant computational load.

2.2. Agricultural systems modeling

APSIM simulates biophysical processes in farming systems (Keating et al., 2003). We used APSIM version 7.3, including its Wheat, SoilN, SoilWater, and Surface Organic Matter (SurfaceOM) modules to model agricultural systems in our study area. The APSIM-Wheat module simulates the growth and development of a wheat crop on a daily time step in response to weather (radiation, temperature), soil water, nitrogen, and residue and crop management. The modules of SurfaceOM and SoilN simulate the decomposition of surface organic matter and the dynamics of carbon