ECOLOGICAL MODELLING 213 (2008) 165–179

available at www.sciencedirect.com

ScienceDirect

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

# A process-based model designed for filling of large data gaps in tower-based measurements of net ecosystem productivity

*Zisheng Xing*[a,e], *Charles P.-A. Bourque*[a,b,*], *Fan-Rui Meng*[a], *Roger M. Cox*[c], *D. Edwin Swift*[c], *Tianshan Zha*[d], *Lien Chow*[e]

[a] *Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, New Brunswick, Canada E3B 6C2*
[b] *Lanzhou Regional Climate Centre, 2070 Donggang East Road, Lanzhou, 730020, Gansu, PR China*
[c] *Natural Resources Canada, Canadian Forest Service, Atlantic Forestry Centre, P.O. Box 4000, Fredericton, New Brunswick, Canada E3B 5P7*
[d] *Climate Research Branch, Environment Canada, 11 Innovation RD, Saskatoon, Saskatchewan, Canada S7N 3H5*
[e] *Potato Research Centre, 850 Lincoln Road, P.O. Box 20280, Fredericton, New Brunswick, Canada E3B 4Z7*

ARTICLE INFO

*Article history:*
Received 20 March 2007
Received in revised form
25 November 2007
Accepted 29 November 2007
Published on line 16 January 2008

*Keywords:*
$CO_2$ flux
Gap-filling model (GFM) for large data gaps
Extrapolation experiments
Net ecosystem productivity
Punch hole experiments

ABSTRACT

In this paper we present a simple hybrid gap-filling model (GFM) designed with a minimum number of parameters necessary to capture the ecological processes important for filling medium-to-large gaps in Flux data. As the model is process-based, the model has potential to be used in filling large gaps exhibiting a broad range of micro-meteorological and site conditions. The GFM performance was evaluated using "*Punch hole*" and extrapolation experiments based on data collected in west-central New Brunswick. These experiments indicated that the GFM is able to provide acceptable results ($r^2 > 0.80$) when >500 data points are used in model parameterization. The GFM was shown to address daytime evolution of NEP reasonably well for a wide range of weather and site conditions. An analysis of residuals indicated that for the most part no obvious trends were evident; although a slight bias was detected in NEP with soil temperature. To explore the portability of the GFM across ecosystem types, a transcontinental validation was conducted using NEP and ancillary data from seven ecosystems along a north-south transect (i.e., temperature–moisture gradient) from northern Europe (Finland) to the Middle East (Israel). The GFM was shown to explain over 75% of the variability in NEP measured at most ecosystems, which strongly suggests that the GFM maybe successfully applied to forest ecosystems outside Canada.



## 1. Introduction

Data gaps in flux measurements (in particular, net ecosystem productivity, NEP) are an ever-present challenge for flux researchers (Greco and Baldocchi, 1996; Aubinet et al., 2000). In general, 17–50% of flux observations are reported as missing or rejected (Falge et al., 2001a,b; Hui et al., 2004). The gaps are caused (*i*) during inclement weather, especially during heavy rainfall and during periods of freezing rain; (*ii*) by equipment failures, maintenance shut-downs, and power outages; or (*iii*)

* *Corresponding author at*: Faculty of Forestry and Environmental Management, University of New Brunswick, Fredericton, New Brunswick, Canada E3B 6C2. Tel.: +1 506 453 4930; fax: +1 506 453 3538.
E-mail address: cbourque@unb.ca (C.P.-A. Bourque).

0304-3800/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2007.11.018

by low wind conditions (Xing et al., 2008). Occasional data gaps in the data stream cause problems by (*i*) reducing the value and integrity of the data, (*ii*) making it difficult to estimate annual NEP budgets and to determine sink–source status of ecosystems, and (*iii*) biasing climatic relationships with NEP, reducing the level of validation (Falge et al., 2001a,b; Hui et al., 2004; Amiro et al., 2005).

The data gaps must be filled in order to derive annual trends and to investigate the time-dependency of the derived quantities (i.e., NEP). Poor gap filling may cause misinterpretation of the data and lead to incorrect conclusions. In previous research, Xing et al. (2007a) have summarized the features of an effective gap-filling protocol. In particular, the method should: (*i*) be easy to implement, (*ii*) be able to capture prominent features of the modelled system and address interactions among relevant environmental controls, (*iii*) have limited number of model parameters for auto-parameterization (parameter optimization), and (*iv*) have few inter-correlated parameters so that model parameterization is internally consistent.

Many gap-filling methodologies have been developed over the past decade, each with their strengths and weaknesses. Some of these methods include the application of (*i*) lookup tables (LUT), (*ii*) non-linear regression (NLR), (*iii*) artificial neural networks (ANN), (*iv*) multiple imputation (MI), and (*v*) ecosystem models (EM). LUT use a randomly or systematically chosen value from individual observations that have similar values on other variables (Falge et al., 2001a,b; Greco and Baldocchi, 1996). LUT, typically consider only the effects of light and temperature in their formulation, restricting the portability of the method to regimes of unchanging soil water content. NLR is used to estimate missing values with mathematical functions generated by regressing dependent variables against explanatory (independent) variables. Most often, one or two explanatory variables are used. Such functions, seldom consider interaction between environmental controls, again restricting their application to a small subset of cases.

ANN generates predictive models (networks) of simple processing elements (i.e., neurons) by iterative learning (Aubinet et al., 2000). The approach is based on a black-box application of statistical formulations (rules) with little biophysical-ecological realism. In order to match training data, the network evolves and adopts almost any parameter value in its search for perfect agreement. In doing so, the network produces ecologically meaningless quantities. If not carefully applied, ANN can over fit (over-generalize) the training dataset and produce unreasonable output. Also, ANN can adopt different solution strategies with the same training data, rendering its application very difficult. Moffat et al. (2007) point out, all ANN methods are complex to implement and require smoothing or regularization to ensure good reliability in the annual NEP sums. Smoothing and regularization introduce their own level of uncertainty in the final results. MI provides a set of values through standard statistical techniques (Hui et al., 2004) often difficult to implement in gap filling.

Ideally, EM provide the best overall solution to gap filling, however, their complicated structures and parameterization requirements make EM cumbersome to use. Additionally, because of the presence of many inter-correlated parameters and non-linearities in the models, auto-parameterization is usually difficult because of internal inconsistencies and problems associated with solution divergence. Typically, parameterization of EM is done by trial and error and is very time consuming. As a general rule, as the number of model parameters increases, the statistical stability of the model decreases.

Some parameters may be set only through specific field measurement techniques, including some for which current instrumentation may be inadequate or does not exist. Available models are designed to serve specific applications with specific spatial and temporal resolutions. Also, time series from variables that are hard to fill reflect the outcome of complex biological interactions among different ecosystem components. Simple linear interpolation and other statistical methods may be good for filling small data gaps lasting a few hours, but are less reliable for filling large gaps lasting more than a few days to a few weeks. To address large gaps, a simple model which combines the flexibility of statistical models with the precision of EM is required to best address gaps in NEP flux measurements. This model should technically balance relevant ecological detail with the number of inter-correlated parameters in model structure.

In Moffat et al.'s (2007) publication, a process-based Biosphere Energy-Transfer Hydrology model (BETHY; Knorr and Kattge, 2005) was used to demonstrate the potential of using process-based models in gap filling of very long periodically interrupted time series of flux data, because of their reduced dependence on existing observations. Xing et al. (2007a,2008) provide two versions of a simplified gap-filling model (GFM) which can be automatically parameterized based on a few available measurement episodes. These GFM's are robust with handling small to intermediately-sized gaps by addressing variations (*i*) in light interception, as a result of differences in sun elevation, sky cloudiness, and partitioning of incident photosynthetically active radiation (PAR) into its direct and diffused components (Xing et al., 2007a, with small gaps), and (*ii*) in air and soil temperatures (Xing et al., 2008, with intermediate-size gaps). Although, the GFM's were shown to underestimate in instances of water stress conditions (Xing et al., 2008), consideration of environmental controls of light and air and soil temperatures on NEP has made the methods very practical in dealing with these two gap-size categories, as long as the input data (mostly meteorological in nature) are available. For large gaps, the leading environmental controls should include water conditions which affect plant growth (and NEP) through either diurnal fluctuations in air relative humidity (RH) and/or long-term fluctuations in soil water content.

In this study, our objectives are to: (*i*) develop an ecosystem-based model with the least number of parameters for gap filling of *large data gaps*; and (*ii*) evaluate the performance of the newly developed GFM for selected scenarios by comparing model results to flux measurements collected at a Fluxnet-Canada Research Network (FCRN) site in west-central New Brunswick (NB), Canada, and seven CarboEurope project sites (as a second validation experiment) using "*Punch hole*" (interpolation, Hui et al., 2004) and various "extrapolation" experiments.