Field Crops Research 227 (2018) 79–90

Contents lists available at ScienceDirect

Field Crops Research

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

Towards a national, remote-sensing-based model for predicting field-scale crop yield

Randall J. Donohue[a,b,*], Roger A. Lawes[c], Gonzalo Mata[c], David Gobbett[c], Jackie Ouzman[c]

[a] CSIRO Land and Water, Building 801, Clunies Ross St, Black Mountain ACT 2601, Australia
[b] ARC Centre of Excellence for Climate System Science, Australia
[c] CSIRO Agriculture and Food, Underwood Ave, Floreat, WA 6014, Australia

ARTICLE INFO

Keywords:
Yield modelling
Canola
Wheat
MODIS

ABSTRACT

Existing agricultural grain yield models predict yield at the field scale, or at regional scales (like districts and countries), but not both with consistent accuracy. Here we describe a scalable, satellite-based yield model called C-Crop. It is calibrated locally and so has field-scale accuracy. Its input data can be inferred remotely (namely crop type, foliage cover and air temperature) and so it can be potentially applied at any regional scale. We calibrated C-Crop using harvester-derived yield data for canola (31 field-years) and wheat (160 field-years), across the Australian cropping zone. C-Crop explained 69 and 68% of the observed variability in field-scale canola and wheat yields, respectively, with errors in the order of 33% and 32% of total yield. Given its simplicity, C-Crop is an effective model for estimating field-scale crop yields and has the potential to be applied across large regions.

1. Introduction

Information about crop yield is often sought at the national scale for food security, grain marketing and grain handling purposes (Becker-Reshef et al., 2010; Wu et al., 2015). Similarly, crop yield information is also sought at the field scale by farmers, advisors and consultants to assist with farm-level management actions (e.g., Raun et al., 2008; Hochman et al., 2009; Lawes and Robertson, 2011; Pahlmann et al., 2017). Unfortunately, the methods used to derive national estimates of grain yield production (e.g., Wu et al., 2015) bear little resemblance to those used to estimate yield at the farm scale (e.g., Raun et al., 2008; Hochman et al., 2009). Thus, it becomes difficult to change between these two spatial scales without changing methods. By inference, this means the impact of farm management decisions cannot be scaled up beyond the field without modifying the assessment methods and, vice versa, national assessments of the effects of policy decisions cannot be downscaled to understand local impacts.

The ideal crop model would predict and forecast field-scale yields across large areas such as districts, states and countries. Recent advances in computing processing power mean point-based models can be employed to generate vast ensembles of simulations over large areas (Lobell et al., 2015; Hochman et al., 2016). However, extending such simulations across large areas introduces substantial challenges for model parameterisation. Different approaches to dealing with this challenge have been demonstrated in the Australian cropping environment, including up-scaling from a limited number of broadly representative simulation sites (Gobbett et al., 2016) or, conversely, applying general parameterisation rules to large numbers of simulations (e.g., Hochman et al., 2016).

Crop yield models can be parameterised locally or regionally. Locally parameterised models require detailed, site-specific observations as inputs and for parameterisation. These can be thought of as *data intensive* models. Their relevance is usually restricted to small areas by their input data requirements. By contrast, regionally parameterised, or *data extensive,* models use pre-existing, more generalised (and what are usually freely available) input data that cover large spatial areas, such as grids of meteorological variables and satellite imagery. Hence, the scale of parameterisation determines the extent over which a model can be applied. To predict yields across a large number of fields, the model must be parameterised regionally (that is, be data extensive).

Somewhat similarly, the scale of model calibration determines the scale at which a model can accurately predict yields. A locally calibrated model is calibrated at field scales and will most accurately predict yield at that scale. To do this, the model requires field-scale observations to calibrate to. Likewise, a regionally calibrated model is calibrated using regional (e.g., districts, states, countries, etc.) statistics and provides best yield estimates summarised by region (e.g., Potgieter et al., 2005). Usually, but not always, a model's parameterisation and

* Corresponding author at: CSIRO Land and Water, Building 801, Clunies Ross St, Black Mountain ACT 2601, Australia.
E-mail address: Randall.Donohue@csiro.au (R.J. Donohue).

https://doi.org/10.1016/j.fcr.2018.08.005
Received 7 February 2018; Received in revised form 7 August 2018; Accepted 7 August 2018

calibration scales go hand-in-hand (locally calibrated *and* locally parameterised, and vice versa) (Strand, 1981; Lobell, 2013). To date this has meant that field-scale accuracy and regional coverage have been largely mutually exclusive model characteristics. Due to either input data restrictions, or limitations to the transferability of calibrations, local models cannot be reliably scaled-up nor can regional models be scaled-down. The challenge in national, field-scale yield prediction, then, is to develop a model that is regionally parameterised (data extensive) but locally calibrated.

Lobell et al. (2015) have recently been progressing such a modelling system. The aim of their study was to scale-up a data-intensive crop model (APSIM) so it could be used to predict field-scale maize yields across regions (three mid-western US states). To overcome the lack of site-specific parameters across the region, Lobell et al. (2015) ran ensembles of APSIM at 6 representative sites, with the ensembles encompassing a wide range of possible conditions relating to fertilisation rates, sowing density, sowing date, cultivar and soil moisture. The APSIM-modelled LAI ensembles were converted to greenness values that are equivalent to those estimated from Landsat imagery. For each possible image date or combination of dates, a linear model was developed from the ensemble of simulations that related meteorological conditions and APSIM greenness to the simulated yields. Then, for each pixel (which have a resolution of 30 m) in the Landsat imagery, and using only the image dates with the highest quality data (i.e., least cloud-affected), the authors applied the linear model to estimate yields from the Landsat greenness values. Lobell et al. (2015)'s model r^2 was 0.35 for field-scale maize yields (with all years pooled; for individual state-year combinations, this ranged between 0.14 to 0.58). For soybean, average r^2 was 0.32 (ranging between 0.03 and 0.55). At the county-scale, the model r^2 increased to 55% of observed maize yields, with an error (RMSE) of about 3.4 t ha^{-1} (Jin et al., 2017). On yields of around 10 t ha^{-1}, this is a relative error of approximately 34%. Later modifications to the model (namely using Landsat to calibrate APSIM phenology and using biomass instead of yield in the linear models) improved the r^2 to 0.75 with an RMSE of 1.8 t ha^{-1} (Jin et al., 2017).

Any approach that seeks to scale-up a data intensive model to operate as a data extensive one will be informationally and computationally demanding. It requires all the intensive (field-scale) input data in the model calibration phase but now requires such data across much larger areas than when applied across just a few fields. Whilst these represent challenges to this approach rather than barriers, our motivation here is to test whether the opposite approach – down-scaling a data extensive model by calibrating it locally – might be just as effective but with greater simplicity.

The generation and testing of a locally calibrated, data-extensive crop model has traditionally been hindered by the lack of wide-extent, field-scale yield data (Gobbett et al., 2016). However, models are now taking advantage of a new source of yield data, which is derived from grain yield harvesters equipped with yield monitoring equipment (e.g., Lawes et al., 2009). This technology has facilitated a rapid increase in the volume of field-based yield observations, and this is changing the prospects of what scales yield models can be calibrated at. Such data are now available for model parameterisation and, for the first time (at least in Australia), there are sufficient data available to generate and test a field-scale, national crop yield model.

Our primary aim here is to develop and test a field-scale, crop yield model that can potentially be applied across Australia. A secondary aim is to use the simplest-possible approach to both lower computational overheads and to increase transparency of model processes. In the following, we describe 'C-Crop', a simple, remote-sensing-based crop yield model. This is a combination of two reasonably standard models – a light use efficiency carbon assimilation model and a carbon accumulation and turn-over model. We calibrate and then validate C-Crop using an extensive dataset of field-scale yield maps spanning the full extent of the Australian cropping zone. We then test a number of variations in the model structure and parameterisation in order to understand some of the model characteristics.

2. Methods

2.1. Overview

C-Crop predicts grain yield, or mass, from remotely sensed greenness data and a carbon mass accumulation and turn-over model. It first predicts total assimilation rates (or gross primary productivity), then above-ground plant carbon (C) mass and then grain yield. The model is applied here to individual fields and at a 16-day time-step. Below we outline the four input datasets required by the model – remotely sensed plant-absorbed photosynthetically active radiation, total solar irradiance, diffuse solar irradiation and air temperature. We then describe the model itself, model calibration and validation, and finally the observational data used for the calibration and validation. We estimate the yield for two of Australia's most important broadacre crops – canola and wheat.

2.2. Input data

There are four input datasets needed for the C-Crop modelling (Table 1). The first input is remotely sensed f_{PAR} data (the fraction of photosynthetically active radiation [i.e., PAR] absorbed by plants; 0.0–1.0). This is derived from the 250 m resolution, collection 5, MODIS 'MOD13Q1' Normalised Difference Vegetation Index data (Justice et al., 1998), which are converted to f_{PAR} using the method of Donohue et al. (2014). This method linearly rescales NDVI to f_{PAR} using scaling thresholds that represents the NDVI of bare soil (no foliage cover) and of full cover. To minimise the effects of background soil colour (Huete and Jackson, 1988; Montandon and Small, 2008), the bare soil NDVI value was obtained separately for each pixel (or grid cell) by identifying the minimum value of each cell's full 15-year (2001–2015) NDVI time-series. This assumes that, at each cell location, at least once over the past 15 years, there was one period when foliage cover was zero. The maximum NDVI value was taken as the maximum value present in all 15-year time-series of all crop types. Hence, there was one minimum NDVI threshold for each cell, and one 'global' maximum NDVI threshold for all cells. This method of calculating site-specific f_{PAR} input data (particularly with respect to soil background effects) has been shown to generate more accurate monthly estimates of GPP than the MOD15A2 f_{PAR} product – as well as the MOD17A2 GPP product itself – when compared to Australian flux tower estimates of GPP (Donohue et al., 2014). An f_{PAR} time-series for each field was obtained by averaging the values of all input 250 m cells that were wholly (100%) contained within each field boundary.

The second input is average daily incoming solar (short-wave) surface irradiance (R_s, MJ m^{-2} d^{-1}). This was calculated at 250 m according to the method of Donohue et al. (2010), which attenuates top-of-atmosphere irradiance according atmospheric depth and an Australia-specific, temperature-based transmissivity function (McVicar and Jupp, 1999). The third input is the diffuse fraction (f_D), which is the ratio of diffuse to total solar irradiance and varies from 0.2 under clear

Table 1
Input data used in C-Crop modelling. All input spatial data have a 250 m resolution.

Variable	Description	Units	Source
f_{PAR}	Fractional absorption of photosynthetically active radiation	0 – 0.95	(Donohue et al., 2014)
R_s	Daily solar irradiance	MJ m^{-2} d^{-1}	(Donohue et al., 2014)
f_D	Fraction of diffuse radiation	0 – 1	(Roderick, 1999)
T	Period-average daily air temperature	°C	(Jones et al., 2009)