Remote Sensing of Environment 138 (2013) 149–164

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction☆

Amor V.M. Ines [a,*], Narendra N. Das [b], James W. Hansen [a], Eni G. Njoku [b]

[a] International Research Institute for Climate and Society, The Earth Institute at Columbia University, 61 Route 9W Palisades, NY 10964, United States
[b] NASA, Jet Propulsion Laboratory, M/S 300-323, 4800 Oak Grove Drive, Pasadena, CA 91109, United States

ARTICLE INFO

Article history:
Received 10 November 2012
Received in revised form 28 June 2013
Accepted 1 July 2013
Available online 15 August 2013

Keywords:
Yield forecasting
Crop model
Data assimilation
Soil moisture
Leaf area index

ABSTRACT

To improve the prediction of crop yields at an aggregate scale, we developed a data assimilation-crop modeling framework that incorporates remotely sensed soil moisture and leaf area index (LAI) into a crop model using sequential data assimilation. The core of the framework is an Ensemble Kalman Filter (EnKF) used to control crop model runs, assimilate remote sensing (RS) data and update model state variables. We modified the Decision Support System for Agro-technology Transfer – Cropping System Model (DSSAT-CSM)-Maize model (Jones et al., 2003) to be able to stop and start simulations at any given time in the growing season, such that the EnKF can update model state variables as RS data become available. The data assimilation-crop modeling framework was evaluated against 2003–2009 maize yields in Story County, Iowa, USA, assimilating AMSR-E soil moisture and MODIS-LAI data independently and simultaneously. Assimilating LAI or soil moisture independently slightly improved the correlation of observed and simulated yields ($R = 0.51$ and 0.50) compared to no data assimilation (open-loop; $R = 0.47$) but prediction errors improved with reductions in MBE and RMSE by 0.5 and 0.5 Mg ha^{-1} respectively for LAI assimilation while these were reduced by 1.8 and 1.1 Mg ha^{-1} for soil moisture assimilation. Yield correlation improved more when both soil moisture and LAI were assimilated ($R = 0.65$) suggesting a cause–effect interaction between soil moisture and LAI, prediction errors (MBE and RMSE) were also reduced by 1.7 and 1.8 Mg ha^{-1} with respect to open-loop simulations. Results suggest that assimilation of LAI independently might be preferable when conditions are extremely wet while assimilation of soil moisture + LAI might be more suitable when conditions are more nominal. AMSR-E soil moisture tends to be more biased under the presence of high vegetation (i.e., when crops are fully developed) and that updating rootzone soil moisture by near-surface soil moisture assimilation under very wet conditions could increase the modeled percolation causing excessive nitrogen (N) leaching hence reducing crop yields even with water stress reduced at a minimum due to soil moisture assimilation. However, applying the data assimilation-crop modeling framework strategically by considering a-priori information on climate condition expected during the growing season may improve yield prediction performance substantially, in our case with higher correlation ($R = 0.80$) and more reductions in MBE and RMSE (2.5 and 3.3 Mg ha^{-1}) compared to when there is no data assimilation. Scaling AMSR-E soil moisture to the climatology of the model did not improve our data assimilation results because the model is also biased. Better soil moisture products e.g., from Soil Moisture Active Passive (SMAP) mission, may solve the soil moisture data issue in the near future.

1. Introduction

When a crop model is used to predict crop yields early in the growing season, two sources of uncertainties prevail — those coming from climate and model uncertainties (Hansen, Challinor, Ines, Wheeler, & Moron, 2006). Climate-related uncertainty is greatest early in the growing season but tends to decrease as weather data become available as growing season progresses. Model-related uncertainty due to errors in model structure, modeling assumptions and other ancillary data, generally remains constant through the growing season. Skillful climate forecasts can reduce climate-related uncertainty in crop yield prediction especially at the earlier stages of the growing season, while model-related uncertainty can potentially be reduced by assimilating remote sensing (RS) data during the growing season (de Wit & Van Diepen, 2007; Hansen et al., 2006; Vazifedoust, Van Dam, Bastiaanssen, & Feddes, 2009).

Remote sensing had been incorporated into crop simulation models either as a forcing function or simulation steering (Bouman, Van Diepen, Vossen, & Van Der Val, 1997). Forcing function is applied to replace simulated state variable with the RS observation while simulation

* Corresponding author. Tel.: +1 8456804490.
E-mail address: ines@iri.columbia.edu (A.V.M. Ines).

0034-4257/$ – see front matter © 2013 The Authors. Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.rse.2013.07.018

steering is used to re-initialize (e.g., sowing date, planting density) or re-parameterize (e.g., canopy and growth parameters) the crop model in a way that minimizes the difference between simulated and measured data. Examples of the simulation steering approach include the works of Bouman (1992), Olioso et al. (2005), Fang, Liang, and Hoogenboom (2011) and Thorp et al. (2012) who linked radiative transfer models with crop models. Ines, Honda, Gupta, Droogers, and Clemente (2006) used remotely sensed evapotranspiration to re-parameterize soil properties, crop and water management parameters of a pseudo-regional Soil–Water–Atmosphere–Plant (SWAP) model.

When RS data are used to replace the value of a model-simulated state variable or to infer some soil–plant–atmosphere–continuum properties, one assumes that the RS data are free of error or assumes that the level of data error is acceptable to be propagated within the simulated system (Fang, Liang, Hoogenboom, Teasdale, & Cavigelli, 2008; Ines & Mohanty, 2008a,b,c; Ines & Mohanty, 2009). Thorp, Hunsaker, and French (2010) assimilated measured Leaf Area Index (LAI) in the DSSAT-CSM-Wheat model using forcing and updating mechanisms. The updating mechanism is a forcing scheme that accounts for back propagation of the change in LAI to the system. Their simple assimilation procedure is more successful in minimizing errors in ET and canopy weight, but had difficulty improving yield simulations because yield is controlled by other factors aside from LAI. Vazifedoust et al. (2009) conducted a simple sequential data assimilation using a constant gain Kalman filter to assimilate LAI and the ratio of actual ET to potential ET (ET/ET_p) in SWAP-WOFOST and found significant improvements in simulated total dry matter but only one among three of their observation fields showed significant improvements in simulated yield. It should be noted that timing of the use and frequency of LAI data assimilation in crop model is critical as LAI (or NDVI) is more directly related to yield at silking and grain filling (Ozalkan, Sepetoglu, Daur, & Sen, 2010; Sehgal, Sastri, Karla, & Dadhwal, 2005; Teal et al., 2006).

Sequential data assimilation is a robust way of combining model and observations to minimize the uncertainty of a given modeled state as it enhances the use of information between imperfect model and observations. Of the several algorithms (e.g., particle filter, Kalman filter) capable of performing data assimilation to update sequentially model states and parameters, the Monte Carlo-based Ensemble Kalman Filter (EnKF) is the one that is widely used (Evensen, 2003). EnKF received a lot of attention in the geosciences because of its ease of implementation, computational efficiency and optimum performance. It uses the Monte Carlo approach to approximate the conditional second-order moments of variables of interest using a finite number of randomly generated model replicates, then corrects model forecast and error covariance (Evensen, 2003; Houtekamer & Mitchell, 1998). Many studies have implemented EnKF to assimilate RS data in meteorological and hydrological models with considerable success (e.g., Crow & Wood, 2003; Das & Mohanty, 2006; Das, Mohanty, Cosh, & Jackson, 2008; Dunne & Entekhabi, 2005; Evensen, 2003; Keppenne & Rienecker, 2002; Reichle, McLaughlin, & Entekhabi, 2002).

As with any other models, crop simulation models are also subject to structural and data (input and forcing) errors hence they are imperfect in simulating the truth. Sequential data assimilation can be used to improve crop model performance without altering its structure by periodically updating state variables within the growing season with RS observations. RS of vegetation (e.g., LAI) and soil moisture are potentially useful for sequential data assimilation because of their obvious influence on crop growth, hence on crop yields. Their spatial and temporal coverage also allows data assimilation for crop forecasting at regional scale.

EnKF had been used with crop models recently with some success and challenges especially when assimilating LAI (e.g., Curnel, de Wit, Duveiller, & Defourny, 2011). Most of these studies however were conducted under hypothetical conditions, so-called forward–backward simulations, and could be limited to explaining fully the strengths and limitations of the method under actual conditions, especially at predicting yield at aggregate scale (Curnel et al., 2011; Nearing et al., 2012). de Wit and Van Diepen (2007) showed the utility of RS-derived rootzone soil wetness index to correct some of the errors in the soil water balance associated with imperfect model inputs e.g., gridded rainfall data, in crop yield prediction.

In this paper, we developed a data assimilation-crop modeling framework for assimilating remotely sensed data with a crop model that could be used to improve crop yield forecasting at a given lead-time within the growing season. We present our implementation of an EnKF data assimilation system, development of the stand-alone DSSAT-CSM-Maize model, and testing and evaluation of the method under actual growing conditions in Story County, Iowa. The testing and evaluation aims to quantify the use of remotely sensed soil moisture and LAI to improve simulated yields within the data assimilation-crop modeling framework, independently and simultaneously. A variant of EnKF called an Ensemble Square Root Filter (Whitaker & Hamill, 2002) (but we termed it EnKF in general) was implemented for this study to simplify the use of RS data in the data assimilation, especially crop growth observations e.g., LAI, as the square root filter allows data assimilation without perturbing the observed data. This kind of work is important to improving the applications of data assimilation in crop yield forecasting.

2. Methods

2.1. Development of EnKF-DSSAT-CSM-Maize

2.1.1. EnKF data assimilation system

The core of data assimilation lies in the Kalman filter system, which assumes that observations are related to the true state x_t (e.g., soil moisture or LAI at time t) as:

$$y = Hx_t + \varepsilon \tag{1}$$

where y is the observation vector, ε is a Gaussian random error vector with a mean of zero and observation error covariance R, and H is the operator that maps the model variable space to the observation space. Furthermore, the forecast of x_t at $t = k$ is Gaussian with mean $x^f_{t=k}$ and error covariance $P^f_{t=k}$. Under these assumptions, the estimated state and error covariance is updated as:

$$x^a_{t=k} = x^f_{t-k} + K\left(y - Hx^f_{t=k}\right) \tag{2}$$

$$P^a_{t=k} = (I - KH)P^f_{t=k} \tag{3}$$

where f and a are indices of the prior (called forecast) and posterior (called analysis) estimates, respectively, t is an index of time, I is the identity matrix, and K is the Kalman gain matrix defined as

$$K = P^f_{t=k}H^T\left(HP^f_{t=k}H^T + R\right)^{-1}\ . \tag{4}$$

The EnKF forecast and analysis error covariance come directly from an ensemble of model simulations:

$$P^fH^T = (N_e - 1)^{-1}\sum_{n=1}^{N_e}\left(x^f_n - \bar{x}^f\right)\left(Hx^f_n - H\bar{x}^f\right)^T \tag{5}$$

where N_e is the number of ensemble members, n is a running index for ensemble member, and $\bar{x}^f$ represents the ensemble mean calculated as:

$$\bar{x}^f = N_e^{-1}\sum_{n=1}^{N_e}x^f_n\ . \tag{6}$$

Usually, the ensemble is generated by perturbing the observed data. The variance used in the perturbation is based on the uncertainty of the