Geosci. Model Dev. Discuss.,
doi:10.5194/gmd-2016-243-RC1, 2016

Geoscientific Model Development Discussions

Interactive comment on "The impacts of data constraints on the predictive performance of a general process-based crop model (PeakN-crop v1.0)" *by* Silvia Caldararu et al.

D. Wallach (Referee)

Daniel.Wallach@toulouse.inra.fr

Received and published: 23 November 2016

General comments

This paper proposes and tests a new approach to estimation of crop yields applicable to sites with available remote sensing data, tower eddy covariance data and regional yield. The overall framework is the quest for a prediction method that is applicable at many or ultimately most sites around the world, a very ambitious goal. The authors correctly argue that the availability of data is the major constraint, and here they innovatively combine data sets which are available at many worldwide sites. A second challenge is to develop a model that can use the available data as inputs for prediction. The authors argue that process based models have major advantages compared to

statistical models for extrapolating to weather conditions that might be experienced under climate change, but that current process based models are not sufficiently general for worldwide use. They therefore propose a new, simple process based model which is supposed to have sufficient generality. The model is based on assuming that plants optimize the root:shoot ratio and also the time of flowering. Finally, the authors test their approach across 15 sites and two crops, wheat and maize.

I think that this is an important paper, because of the issues raised and the innovativeness of the proposed solutions. There are many possible criticisms, and I detail a certain number below. That is, the authors have not found a satisfactory answer to the overall problem of predicting yields at arbitrary global locations. That is hardly surprising given the magnitude of the objective, and should not obscure the real contributions of the study.

Specific comments

A first criticism is that there seems to be some ambiguity about the exact objective of the predictions. In the introduction the authors speak of application to a "generic farm location", whereas evaluation is based on comparison with regional yields. Prediction for a farm, with uniform management, is quite different than prediction for a region. The paper seems more oriented toward regional prediction, since county yields are one of the data sets used as input and are also used for evaluation. On the other hand, the landcover input data was aggregated to 3km by 3 km pixels, which is generally intermediate between farm and county scale. In any case, it is essential to clarify the spatial scale of interest.

Much of the evaluation is based on comparing the model using data constrained parameters to the model with prior parameters. This is not a very interesting comparison. The prior parameters were chosen quite arbitrarily by the authors to represent essentially a total lack of information about the parameter values. The fact that adding some information improves the situation is hardly surprising. A much more relevant com-

parison would be between the data constrained model and long term average county yields. Does the model do better than simply assuming that the future is like the average of the past? This is analogous to comparing climate forecasts with climatology.

Another aspect of the evaluation is that uncertainty intervals are given for each prediction. This is extremely informative and pertinent, and is a very valuable addition to the comparison between the mean prediction and observed values. However, the uncertainty results need to be discussed more thoroughly. For example, it seems that the 95% intervals for yield cover all historic yields at most sites (Fig. 3a). Surely this uncertainty is so large as to render the results useless. More discussion is required here.

The uncertainty calculations are based on propagating uncertainty in the parameter values through the model. It is not clear if residual error is included when calculating uncertainty intervals or not. It should be. Also, there are other sources of uncertainty than the parameters which might be quite important, in particular uncertainty about management practices. This should at least be discussed.

It seems that the likelihood used here for the Bayesian estimation assumes that all data are independent. This is of course almost certainly false for time series data. Taking non independence into account by dividing by the number of measurements is only a very crude approximation.

The model that is proposed is an original model, based on the assumption that plants optimize partitioning between roots and aboveground biomass, as well as time of flowering. The major advantage of such a model is that it allows the same model, with the same parameters, to be used for different cultivars of the same species, if one accepts that the cultivars chosen for a particular location are optimized for the environment there. More detail about the model would be helpful. How exactly is the date of flowering calculated? According to the text, the switch from vegetative to reproductive growth occurs when increased vegetative fractions would not result in an overall increase in

growth rate. Is this calculated day by day or is there some averaging over environmental conditions to ensure that the plant doesn't respond to conditions on one specific day? What exactly are the management inputs required for the model? The authors mention sowing and harvest date, but aren't sowing density and fertilizer inputs also required? The required management information should be made clear, as well as the sources of this information.

There is also no information on soils. Apparently this information is not needed here thanks to the assumptions that there is no water limitation, and that initial soil N is negligible compared to fertilizer N. In general, however, it will be necessary to have soil information.

The authors suggest that the model could be tested by comparing different model structures. Perhaps more useful would be to test the model proposed here with much more detailed input data, in order to reduce the data as a source of error and thereby isolate the amount of error due to the model. Technical corrections

P7 L23-24. "given the model" needs to be omitted

Interactive comment on Geosci. Model Dev. Discuss., doi:10.5194/gmd-2016-243, 2016.