In-season probabilistic crop yield forecasting, integrating agro-climate, remote-sensing and phenology data

Nathaniel K. Newlands[1,2], David S. Zamar[1]
[1]Science and Technology Branch, Agriculture and Agri-Food Canada,
Lethbridge Research Centre, 5403 -1st Ave. S., P.O. Box 3000,
Lethbridge, Alberta, Canada, T1J 4B1
[2]Department of Statistics, University of British Columbia,
Earth Sciences Building, 2207 Main Mall, Vancouver, British Columbia, V6T 1Z4

Abstract
Statistical models help to provide decision-makers with an improved ability to spatially identify and assess, with enhanced foresight, potential risks and vulnerability of natural resources to climate variability and extremes. They also enable integration of diverse geospatial information together with its uncertainty for operational real-world application. We showcase a Bayesian method for sequential forecasting of the yield of major crops grown across the Canadian Prairies, Western Canada. This method incorporates robust least angle regression followed by robust cross validation for variable-selection, Markov chain Monte Carlo (MCMC) sampling for added spatial correlation support, and forms a joint probability distribution using the random forests algorithm for non-parametric modeling of future observable variables. We explore the relative improvement of candidate agro-climate, remote-sensing, and phenology indices on the overall accuracy of in-season forecasts (updated on a monthly basis) at two different spatial resolutions. Preliminary findings from cross-validation on spring wheat yield indicate a gain of 10% when involving net-difference vegetation index (NDVI) as a spatial index of crop yield potential and net model accuracy of 89%.

Key Words: Agriculture, Bayesian, Climate, Forecasting, Risk, Uncertainty

1. Introduction

Canada is a major supplier in the global market of wheat with roughly 70% of its production being exported annually (i.e., includes hard red winter, hard red spring, white wheat and durum varieties). Global wheat production decreased in 2010-11 by 35 million tonnes (Mt) from 647.6 Mt in 2009-10. Similarly, Canadian wheat production decreased by 1.3 Mt from 2009-2010 to 20.1 Mt, mainly attributed to 6% lower seeded area and lower yields as a result of excess precipitation/soil moisture, a cool summer, with wet weather through the harvest period in most growing areas (2011 Market Outlook Report, Statistics Canada and Agriculture and Agri-Food Canada/AAFC). Such changes exemplify just how much weather during the growing season, alongside long-term changes in climate, are having a significant impact on regional and global wheat production.

While 2000-04 was an unprecedented drought period in North America, this year (i.e., 2012), a total of 64% of land across the Great Plains/Midwestern United States is under abnormally dry to exceptional drought conditions (United States Department of

Agriculture (USDA) Drought Monitor, http://droughtmonitor.unl.edu/). The National Oceanic and Atmospheric Administration (NOAA) reports that January-June this year was the warmest first half of *any* year on record for the contiguous United States, and the past twelve months have been the warmest the United States has experienced since the dawn of record-keeping in 1895 (National Climatic Data Center). This year's drought is the most serious to impact U.S. agriculture since 1988. Nearly 50% of corn and 37 % of the soybeans grown in the US were rated poor to very poor, with three-quarters of U.S. cattle acreage in drought-affected areas. The 1988 drought delivered a $77 billion loss due to crop, landscape, and other damages (USDA's National Agricultural Statistics Service). Drought also continues to impact other food-producing regions in Russia, China, North/South Korea, raising the prospect of higher commodity prices and localized food shortages (Global Information and Early-Warning System (GIEWS), Food and Agriculture Organization of the United Nations/FAO). As more extreme weather events are anticipated to accompany a warming climate, improved methods for forecasting crop production and its response to climate and other agronomic factors, is becoming increasingly important to guide agricultural producers in making more informed in-season crop management and financial decisions. The broader agricultural industry and government policy-makers increasingly rely on crop market outlooks and yield forecasts for their decision-making.

Given the broad environmental and monetary impact of extreme weather events, a crop forecasting methodology/operational system that can integrate best available data to accurately and sequentially forecast in-season or longer-term crop production (i.e., cropped area and yields) offers invaluable benefits to ensuring a robust food supply and enhanced global food security. Crop prediction or forecasting can provide an objective scientific basis for foresighting and assessing the probability of occurrence as well as level of risk associated with future climate changes and other crop-related conditions. Furthermore, by integrating available satellite remote-sensing data having broad spatial coverage and regular updating/repeats over cropland areas of interest, more near-real time information can be obtained, thereby helping to guide more robust and reliable crop forecasts or 'outlooks' within crop growing seasons. Integrating remote-sensing information enables spatial prediction, lessening the importance of observer-based field survey data and enabling them to be more targeted to areas of higher yield variability and associated potential crop production risk.

Increasing societal awareness of climate-driven impacts on crops highlights the scientific importance of generating reliable outlooks of crop yield/production for improved environmental, economic and policy-related decision making. Around the world, many countries have on-going research and development (R&D) activities focusing on improved crop prediction/forecasting (e.g., Australia, Africa (e.g., Zimbabwe), Argentina, China, South Korea, Japan, Philippines, and Canada). In Australia, since 2001, operational, seasonal wheat outlooks have been generated by a regional commodity forecasting system (RCFS) focusing on major cereal crops at the monthly scale for state and shire areas. These forecasts that include global climate model (GCM)-based seasonal rainfall forecasts coupled to a wheat simulation model applied to district and state yields. Such forecasts also include correction for inter-annual phases of El-Nino-Southern Oscillation (ENSO), and the Southern Oscillation Index (SOI) climate oscillation/teleconnections (Hansen *et al.,* 2004). GCM-model climate output has also been used to validate crop forecasts made in Europe using the Sirius wheat simulation model to understand the relative benefit of using dynamical model output relative to historical climatological data (Semenov and Doblas-Reyes, 2007). Since 2004,

AgrometShell (AMS), outlooks for maize, wheat and beans across Eastern Africa are generated 2-3 times over the growing season, primarily relying on indices derived from agro-climate data. In the Philippines, since 2007, an exploratory knowledge-based crop forecasting system focusing on the major crops (i.e., rice, coconuts, bananas, sugarcane) has been developed using monthly remote-sensing, agro-climate and crop phenology data within a Geographical Information System (ArcGIS) (Lansigan *et al.*, 2007). Since 1998, China has run the CropWatch operational system using remote-sensing data (Wu, 2006). In Pakistan, a crop forecasting system has used NDVI and absorbed photosynthetic active radiation (APAR) coupled to radiative energy transfer-balance process algorithm (Bastiaanssen and Ali, 2003). Since 1992, the MARS (Monitoring Agriculture with Remote Sensing) Crop Yield Forecasting System (MCYFS) has been operational providing crop forecasts across the European Continent, including Maghreb and Turkey**,** for wheat, spring barley, grain maize, rape seed, sunflower, potato, sugar beet, field bean, pastures, rice, and soybean) (Joint Research Centre, 2012).

Different levels of engagement relate to available R&D funding, available input data streams, and choice, complexity and extent of innovation of a given model-data assimilation approach (Hansen *et al*., 2006, Stone and Meinke, 2005, Hamner *et al*., 2001). Several major characteristics of crop-climate forecasting which currently pose a considerable challenge relate to: 1) the probabilistic nature of forecasts and associated spatial impacts, 2) the need for forecasts to be assessed in a dynamic framework, 3) high levels of bias and mismatch of spatial and temporal scales between coupled ocean-atmosphere climate- and crop growth- model forecasts, and, 4) whether required institutional and policy arrangements exist to provide a broader environmental-economic agricultural risk framework to ensure crop forecasts offer a beneficial, viable option for producers to improve their economic livelihood and adopt in the longer term. Such arrangements must provide sufficient flexibility in agricultural management to be able to respond to different levels of perceived and real needs and risks (Rubas *et al.*, 2006). De Wit and van Diepen (2007) have explored how probabilistic crop forecasts can be improved with the use of sequential data assimilation techniques such as the Ensemble Kalman filter to integrate soil moisture estimates derived from high-resolution satellite imagery into a probabilistic version of the deterministic, WOFOST (World Food Studies) crop growth-simulation model. While many countries/governments are actively engaged in pursuing operational approaches at a national agricultural 'systems-level', there is parallel efforts to develop and implement farm enterprise-level technological designs to provide enhanced agronomic monitoring 'precision agriculture' (Challinor *et al*., 2003). However, the rapid increase in the availability of remote-sensing data is leading to increased collaborative efforts that focus on identifying efficient methodologies/designs capable of integrating diverse scientific knowledge/findings and data available for cost-effective deployment as real-time, operational forecasting systems. Methodologies or designs that combine models and data to generate crop outlooks are inherently integrative, given that a variety of different sources of climate and environmental data are required to measure and track agricultural/cropland 'productivity', crop condition (i.e., quality), including crop yield (i.e., quantity) or crop production, as mean yield over a given area.

Statistical approaches for modeling crop yield span a broad range – from use of calibrated simplified reference curves/semi-empirical equations, to statistical models, and more complex agro-ecosystem process models. Recently, Budong *et al.*, (2009) applied linear regression to predict spring wheat crop yield at the provincial scale achieving 53-77% accuracy relying solely on agro-climate indices (i.e., growing degree day (GDD),

precipitation (P), actual/potential evapotranspiration (AET, PET) and water stress, SI). Mkhabela *et al.* (2011) have shown that ignoring agro-climate information, and relying on the normalized-difference vegetation remote-sensing index (NDVI) at the CAR scale alone achieves 47-80% accuracy to historical trends. Bornn and Zidek (2012) have combined linear regression with principal component analysis (PCA) multivariate dimensional reduction to GDD and SI indices at the Agricultural Census Region (CAR) scale, achieving 60-70% accuracy. They also modeled the spatial dependence of crop yield between CAR's for these agro-climate variables. Our modeling approach is novel as it extends previous modeling by predicting/forecasting crop yield using statistical MCMC with sequential updating, whereas the method proposed by Bornn & Zidek (2010) is based on estimating crop yield, which requires the realized values of all variables as input. Also, instead of applying PCA, we use model-variable selection techniques to identify the leading predictors of crop yield within each CAR and time-step. Our primary statistical modeling objectives were to: 1) contribute to methodological improvements in crop forecasting by devising and testing a Bayesian/probabilistic approach with sequential-updating to enable input of near-real time information within a crop growing season (i.e., 'in-season' updating), 2) to integrate agro-climate and remote-sensing data in generating crop outlooks, and to, 3) incorporate information from neighboring regions to better represent spatial covariance of crop yield predictors. After further independent validation of our statistical methodology our aim was then to adapt and further refine the method, as required, to ensure it was sufficiently reliable for broader use in generating in-season crop outlooks for the agricultural industry. The longer-term objective of this statistical modeling research work is to provide decision-makers with improved ability to identify and assess potential risks, vulnerability from climate variability/extremes, and to enable the integration of diverse geospatial information together with its uncertainty for operational real-world application. In this paper, we showcase our Bayesian model for integrated crop forecasting, present associated cross-validation results, and introduce a prototype crop outlook for the 2011 growing season (spring wheat across the Canadian Prairies).

2. Methodology

2.1 Study Region and Data Sources

Our study region is shown in Figure 1. Cropland, native grassland (rangeland) and water area is indicated by the color legend. Agricultural Statistical Census Regions (i.e., CAR's) are shown delineated in red (total of 40 spanning the study region) (Statistics Canada, 2007). Green dots show the location of 259 climate stations (i.e., long-term and near-real -time (NRT) monitoring), across Alberta, Saskatchewan and Manitoba Canadian provinces. Only those climate stations distributed in the cropland that have less than 10% missing data were selected. Historical crop yield data from 1976-2011 for each of the CARs was obtained from the Field Crop Reporting Series of the Agriculture Division, Statistics Canada (Statistics Canada, 2012a).

2.1.1 Use of the Normalized-Difference Vegetation Index (NDVI)

Weekly NDVI imagery data was compiled from historical data from the Advanced Very High Resolution Radiometer (AVHRR, ~1km resolution), U.S. National Oceanographic and Atmospheric Administration (NOAA) for years 1987-2011. From 2000 onwards, NDVI data, available from the NOAA Terra satellite's MODerate-resolution Imaging Spectroradiometer (MODIS, 250 m resolution) was utilized.

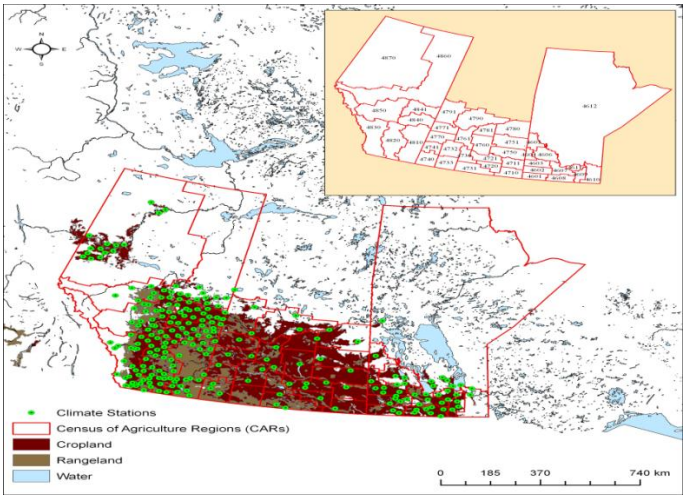

Figure 1. Western Canada/Canadian Prairies (Alberta, Saskatchewan, Manitoba)

Normalized difference vegetation index (NDVI) is derived from bands 1 and 2 of the MODIS detector, and is defined as,

$$NDVI = (\rho_{NIR} - \rho_{RED})/ (\rho_{NIR} + \rho_{RED}), \qquad (1)$$

where ρ_{NIR} and ρ_{RED} are the near-infrared and infrared portions of the electromagnetic spectrum of (0.75-1.5 μm or 841-876 nm) and (0.6-0.7 μm or 620-670 nm), respectively. The NDVI historical time-series were then combined to generate weekly composites south of 60^{o}N (i.e., across the agricultural land area of Canada) for assessing and discriminating crops at the 250m resolution. Besides their use in tracking vegetation/crop growth, NDVI anomalies compare current growing conditions to those in a previous week, the same week in previous years or to a historical mean, and provide useful information for forecasting and assessing drought/floods, crop insurance and tax deferral agricultural support programs.

2.1.1 Use of Agro-climate Indices

Station-based daily temperature and precipitation data are provided by Environment Canada and other partner institutions through a Drought Watch program (http://www.agr.gc.ca/pfra/drought/index_e.htm) operated at the National Agro-climate Information Service (NAIS) of AAFC. This data has been quality controlled and interpolated to provide a continuous, historical daily time-series from 1987-2011. The agro-climatic indices used in the current model include: seeding date (SD), growing degree days above based ambient air temperature of 5^{0}C (GDD), precipitation (P), percent of PAWHC (%PAWHC) and a soil water stress index (SI) defined as SI = 1 - AET/PET, where AET and PET are actual and potential evapotranspiration respectively. To represent the agro-climate of each CAR, average values of all the stations in the cropland of that CAR were calculated for all the agro-climatic indices. Plant Available Soil Water Holding Capacity (PAWHC) at the location of a climate station was determined from soil data obtained from the Canadian Soil Information System (CANSIS). The station-based temperature, precipitation and PAWHC are input into a crop-specific soil water balance calibration called the Versatile Soil Moisture Budget (VSMB) that then generates the set of agro-climatic indices. For some CAR regions there are very few climate stations or the stations are unevenly distributed within the cropland

(e.g. 4751, 4780, 4781, 4790, refer to CAR ID's indicated in the inset of Figure 1). In such instances, station data from neighboring CAR's can be incorporated if a sufficient homogeneity exists. Daily agro-climatic and weekly NDVI indices were temporally-averaged i.e., aggregated into monthly averages and aggregated into three-week moving means, respectively. Figure 2 shows the pattern and amplitude of variability of the input agro-climate and NDVI indices across all CAR's and years.

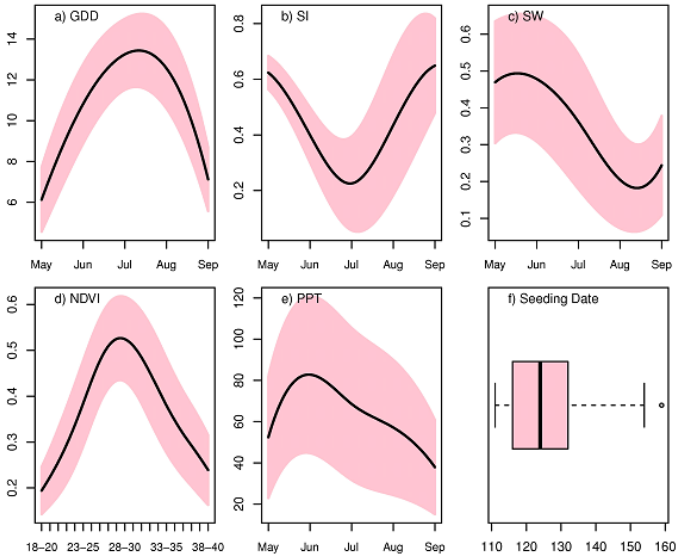

Figure 2. The observed distribution of agro-climate and NDVI remote sensing indices across 1988-2006. Panels (a) to (c) and (e) show the average monthly trend (solid black line) and corresponding 95% confidence band (shaded area) for growing degree days (GDD), stress index (SI), soil water (SW), and precipitation (PPT), respectively. Panel (d) shows the weekly moving average of the normalized vegetation index (NDVI). Panel (f) is a boxplot of the estimated seeding date.

2.2 Bayesian forecasting model

An overview of the modules (i.e., major components) of the spatial, non-parametric Bayesian forecast model is provided in Figure 3. A detailed description of the Bayesian model equations is not presented here, but will be provided in a future publication, as a follow-up to this communication. This Bayesian methodology integrates agro-climate, NDVI remote-sensing indices, and crop phenology data inputs and is applied here in forecasting spring wheat yield at the monthly-scale within Agricultural Census Regions (CAR's) spanning the major agricultural zone of Western Canada (i.e., the Canadian Prairies). The general equation relating crop yield to the agro-climate and NDVI indices within each CAR and three-week moving time-window is:

$$Y_t = \alpha_0 + \alpha_1 t + \sum_{i=2}^{n} \alpha_i X_{t,i} + \varepsilon_t, \qquad (2)$$

where Y_t is the crop yield of year t, α_0 is the regression intercept, $\alpha_1 t$ represent a technology trend of yield over years, $X_{t,i}$ is the predictor i in year t, i is a predictor variable selected from NDVI or agro-climate indices within each three-weeks or month time-segment, and, ε_t is an independent error term with distribution N(0,1).

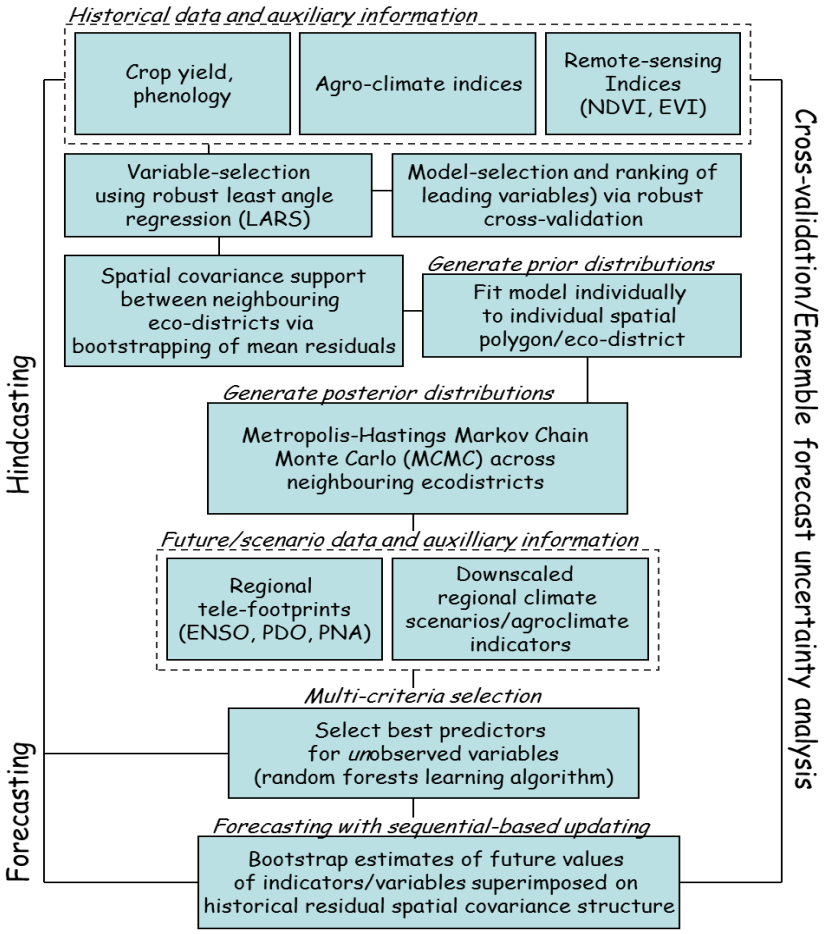

Figure 3. Overview of major modules/components of the Bayesian forecasting model

To account for heteroscedasticity and outliers in the historic data robust regression was used in place of ordinary least squares (OLS) as heteroscedasticity allows the variance to be dependent on the values of the explanatory variables and thus gives more flexibility in modeling the residuals. Robust least angle regression (R-LARS) (using the brlars method) coupled with robust cross validation is used for variable selection (Khan et al., 2007, 2010) (refer to Figure 3). The advantage of using R-LARS is that it is not influenced by the presence of outliers in the data. Thus, variable selection is made for the prediction of non-outlier cases. When selecting the neighbours of a given CAR, the model fit for that CAR is then regressed to available data from all the other CARs. The fitted models from these CARs are then cross-validated using the data for the given CAR. The top *k* ranked CARs, based on the cross-validation error, was selected as neighbours. This method was found to generate a more meaningful prior distribution of the model parameters by providing additional spatial covariance support, as they are coming from CARs in which the corresponding variables were found to be good predictors of crop yield.

Initially, prediction of future values of model variables for a given CAR was made using only the information from the selected model variables. This assumed a Gaussian prior (predictor variable) distribution as a conjugate prior for the joint multivariate Gaussian posterior likelihood distribution. It is well-known that choosing a conjugate prior ensures that the resulting posterior distribution is of the same distribution family as the prior (with

a closed-form solution), and helps to avoid over-fitting on small training samples. The selected predictor variables (for yield) were not good predictors of one another, because they are mostly uncorrelated as seen in our data. Our method therefore uses the entire set of available variables when selecting the best subset of predictors that jointly estimate the unobserved values of future variables. This was accomplished for each CAR using the Random Forests algorithm that creates multiple boot-strapped regression trees without pruning and averages the outputs, and has been found to be very effective in reducing variance and error in high dimensional data sets (Breiman, 2001). Incorporation of the non-parametric Bayesian priors gives us the flexibility to model a wide set of variables and not have to assume a conjugate prior, which in many cases may be inappropriate. Model complexity is automatically determined by the model selection method used (R-LARS followed by robust cross validation) where a maximum model size is initially set by the user.

The module "future scenario data and auxiliary information", whereby downscaled regional climate model (RCM) scenario output and inter-annual climate teleconnection anomalies are also incorporated as indices to help guide the model's in-season forecasts, is discussed as part of future work (see Section 5). The forecasting model was coded and tested using the open source software and statistical libraries provided by the R Statistical Software (R Development Core Team, 2008). The ArcGIS™ (ESRI™, Version 10, 2010) Geographical Information System was used to visualize model output, processing spatial data and generating crop outlook maps.

3. Results

Example sensitivity output of the predictor variable-selection is shown in Figure 4. This plot shows an expected decrease in root-mean-square-error (RMSE) as the number of selected predictor variables increases from 1 to a maximum of 10 variables. Outlier values are indicated as open circles outside of the 2σ confidence intervals. A maximum of five predictors was specified in fitting the historical trends of crop yield within each CAR representing an average level of model complexity and associated cross-validation error.

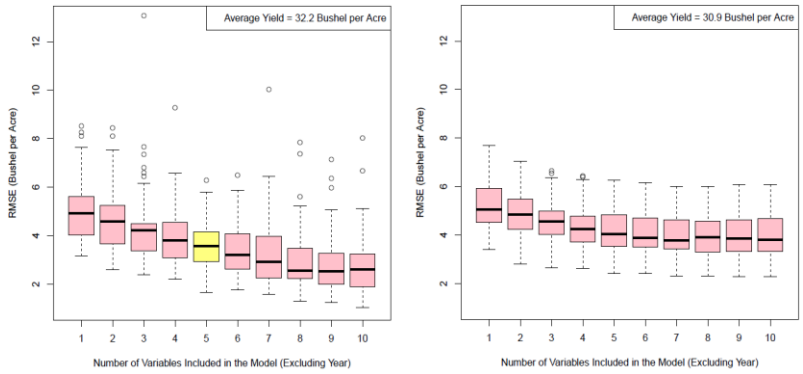

Figure 4. Change of variable-selection cross-validation root-mean square error (RMSE) with increasing number of model variables (model complexity) for a given average crop yield (i.e., 30.9 and 32.2 bushels/acre). Using the NDVI vegetation index, in addition to agro-climate variables that correlate to crop yield (spring wheat), significantly decreases model error.

'Leave-one-out' cross-validation was performed involving a hind-casting of forecasted model output against historical values. The training window used was 1988-2010, whereby a selected year to forecast was omitted. For a given forecast window (i.e., single year), the model's run parameters were: spring-wheat (crop type), 3-week (time-step), scale (40 CAR's), 5 (maximum number of yield predictors or covariates), 3 (number of spatial unit neighbours), 500 (total bootstrap samples used to generate empirical prior distribution), 5000 (chain size for the Markov-Chain Monte-Carlo, MCMC). Computed hind-cast and forecast mean absolute error (MAE) with associated 95% MCMC distribution-based prediction CI's are shown in Figure 5, for each month of the wheat growing season (Canada), and with/without use of remote-sensing data (NDVI index). Outlier CAR's are evident outside of the prediction confidence intervals (i.e., CAR ID#'s 4607, 4609) typically indicating drought/flooding conditions were occurring. Similarly, in the forecast error plots, outlier years are evident that experienced extreme precipitation/temperature conditions (i.e., 1992, 1998, 2006). A summary of these mean model errors and its variability is provided in Table 1. Also, spatial maps of the RMSE error variance are provided in Figure 6 across the Canadian Prairies when including both agro-climate and NDVI indices.

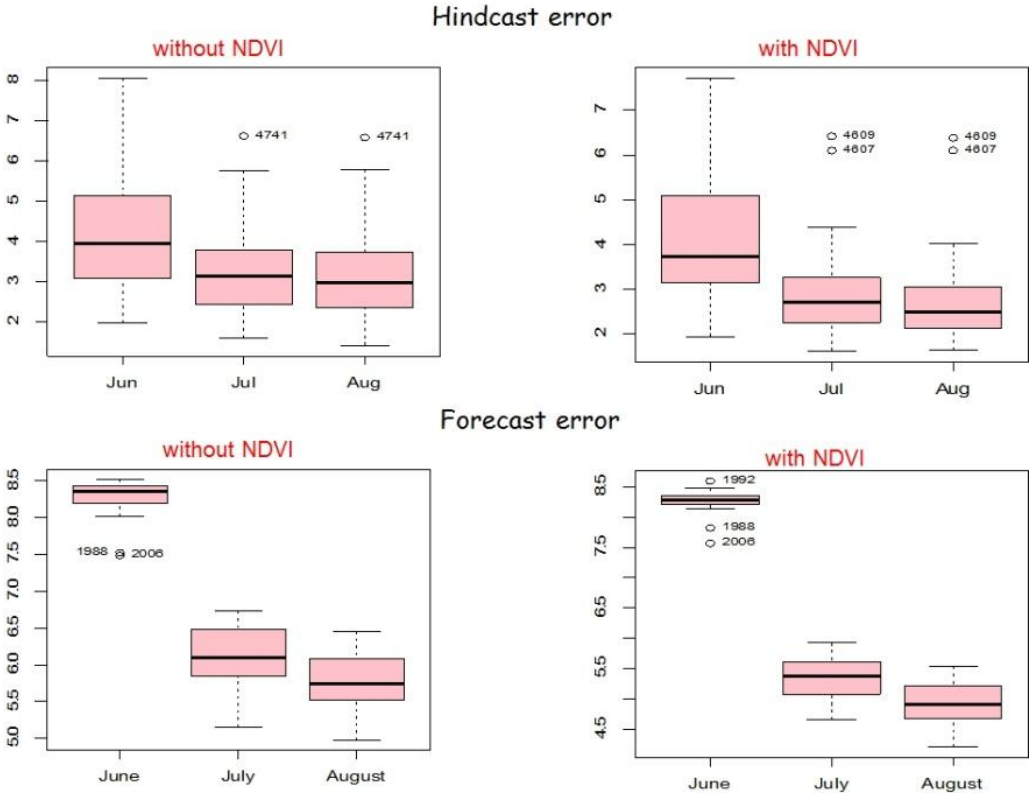

Figure 5. Relative gain in model accuracy when using the NDVI auxiliary index in in-season forecasting crop yield (spring wheat). *Top*: Hindcast error (Mean Absolute Error, MAE) (bushels/acre) obtained from leave-one-out cross-validation against historical crop yield data. *Bottom*: Forecast error based on forecast margin of error (i.e., for a future 3-week time-step).

Table 1: Results from benchmarking of model hindcast and forecast mean absolute error (MAE) under a nested approach (i.e., base agro-climate input data with and without each of the auxiliary vegetation indices, NDVI (1988-2010) derived from MODIS remote-sensing imagery data. Values are for end-of-growing season (i.e., August 1) in units of bushels/acre (bu/ac). The historical reference period used was 1988-2010 for hindcasting with leave-one-out cross-validation (i.e., by removing one year at a time for all spatial units). Forecast error is based on the standard forecast margin of error (i.e., 1 σ). Numbers in brackets are the lower and upper 95% confidence interval, respectively.

Input data	*Hindcast error (CAR-scale)*	*Forecast error (CAR-scale)*
Agro-climate only	*3.0 (1.0, 6.0)*	*5.8 (5.0, 6.8)*
Agro-climate with NDVI	*2.5 (1.0, 4.0)*	*5.0(4.0, 5.5)*

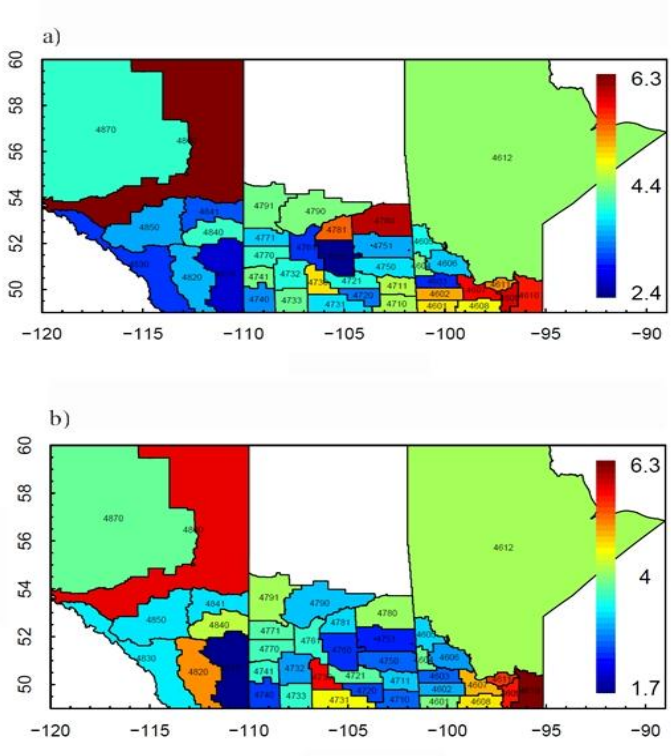

Figure 6. Spatial plots of the error variance (i.e., root mean square error RMSE) obtained when fitting the multivariate crop yield regression model *involving both agro-climate and NDVI vegetation index* to historical spring wheat yield data (1988-2006) across Census of Agriculture (CAR) spatial subdivisions. Inset a) shows model performance based only considering agro-climate variables (refer to variables and their observed/empirical distributions shown in Figure 1), and inset b) shows relative decrease in RMSE error with NDVI vegetation index as an auxiliary variable in addition to agro-climate. Colour legend: units of bushels/acre (bu/ac).

4. Discussion and Conclusions

Our Bayesian forecasting model enables in-season sequential updating, provides automated variable-selection, uses latest non-parametric statistical techniques for forecasting, and models the spatial covariance in yield. Our results show that integrating NDVI reduces cross-validation error by 10%. Hindcast and forecast error values show some improvement (achieves 89% accuracy) compared with forecast model error from the Australian regional crop forecasting model output for the same reference crop (spring wheat). However, large CAR areas with less cropland show high RMSE error, indicating that use of a spatial unit with more homogenous area would be favourable so that spatial areas closely overlap or coincide with actual growing season cropland growing areas for spring wheat. CAR's in southern Manitoba also show higher error, explainable in part due to drought and flooding conditions that have occurred in these regions. Here, additional climate station data and use of additional remote-sensing indices (such as the enhanced vegetation index (EVI) that incorporates the blue portion of the light spectrum would help to better track crop yield by reducing sensitivity of the NDVI index/backscatter to water. Pending further refinements and validation work, this model is anticipated to be generating crop forecasts across Canada as part of an operational version of our prototype model (Canadian Crop Yield Forecaster) or CCYF (i.e., national coverage of major agricultural areas) by 2014-15, delivered by AAFC and publically-released in partnership with Statistic Canada's Crop Condition Assessment Program'.

5. Future Work

We provide here a summary, from a broader perspective, of current work in progress aimed at further enhancing our integrated crop forecasting method: specifically to: 1) integrate crop phenology data and better track important changes in phenology stages (i.e., crucial to accurately tracking climatic response and survival risk linked with their development and growth requirements), 2) integrate auxiliary indices associated with future regional climate model (RCM) scenario output and 3) the spatial impact of regional, inter-annual yield variation driven by climate teleconnections. We aim also to explore delivering crop forecasts/outlook reports to agricultural-related stakeholders, generated using this integrated crop forecasting method, using a range of information and communication technologies (ICT's) that include wireless mobile technology.

5.1 Further understanding of data and model uncertainty

Currently the CCYF prototype is being further validated under different crop types and agricultural regions/climate conditions across Canada. Results from validation against data and at different spatial scales will be important for guiding future improvements across main crop types in Ontario/Prairies. These may indicate that uncertainty can be further reduced by improving the variable selection in the front-end of the modeling methodology; rather than identifying a variable or changing set of climate and EO predictors for each spatial sub-unit, the selection set could instead be *fixed* across all units to a specific set of predictors and only change between each forecast increment or time-step (spatially fixed, temporally variable selection). Further validation work could focus on the algorithms employed and the relative improvement in accuracy when using different/competing statistical algorithms – this is particularly important for the forecasting method's main inference step that currently employs Random Forests Algorithm, but for which other approaches are possible. This benchmarking should be completed before new data is integrated into the methodology (i.e., forecast climate

scenario data). Also, sensitivity analysis of the method's simulation core parameters (crop type, time-step, spatial resolution, total number of possible covariates, number of neighbors, number of bootstrap samples in forming empirical priors, MCMC chain size) could be explored further alongside sensitivity of various operational parameters (NDVI vs. EVI and specific time-windows in the growing season when to integrate various EO/auxiliary indicators).

5.2 Integrating crop phenology information

Survey data from Statistics Canada is currently being used to validate the CCYF prototype for major crops in both the Canadian Prairies and Southern Ontario (i.e., beyond main results obtained so far for Spring Wheat). This requires adapting the method to different crops based on their water, planting and other requirements i.e. a crop phenology component. Currently, crop phenology assumptions are adjusted by the VSMB model that in turn generates agro-climate variables (evapotranspiration, stress index etc..) used as input by the CCYF. However, it would be beneficial to, in the future, to build a crop phenology module that is separate from VSMB itself and link it within the CCYF methodology. This would enable some extended development on crop phenology such as – comparison of adjustments based on several approaches, such as: 1) assuming *fixed* number and width of phenology intervals (EGDD/GDD, water thresholds etc..) using a look-up table (LUT) approach, or, 2) assuming *variable* number of phenology stages and their staging width/length, assuming some degree of water-use efficiency and optimization by the crop itself in adjusting to specific agro-climate thresholds alongside additional crop stress-related requirements. This approach would select the main covariates within each phenology stage, rather than assume all are relevant across all stages, or 3) an approach whereby NDVI, EVI, leaf area index (LAI), SAR-early season soil moisture and fraction of photosynthetic active radiation (FPAR) –related earth observation (EO) indices are used to help identify more precise phenology staging assumptions, incorporated across certain stages of crop phenology that match when they are most accurate (minimal uncertainty linked with backscatter effects).

5.3 Integrating downscaled climate scenario into crop forecast outlooks

The crop forecasts would be considerably enhanced by integrating downscaled future climate from regional climate model scenario (RCM) output. This information is currently being statistical validated at the daily, 10km scale for weekly aggregation required by the CCYF method. Depending on what spatial resolution the method performs best across operationally (from step 1 above). it will be important to downscale future climate data to the same operational scale. This will rely on a method for statistical downscaling. Nonetheless, scenario data across a range of scales should be used to verify the scale with the minimum RMSE variance uncertainty. Based on this, data downscaled to the 25km (from the 50 km typically RCM output scale), for example, may correlate sufficiently well. Inclusion of future climate data will help to fine-tune the method, and make it more robust to forecasting under significant perturbations associated with climate extremes. Including future climate scenario information will greatly enhance the CCYF forecasting capabilities.

5.4 Spatial impacts from interacting teleconnections

AAFC's Science and Technology Branch (S&T) is continuing to improve the analysis of data information as well as developing better methods for understanding and exploring the spatial effects of regionally-relevant teleconnections/oscillations on crop survival and growth (such as the El-Nino Southern Oscillation (ENSO) related to sea surface

temperature changes and Pacific North American Oscillation (PNA) related to geo-potential height across Western Canada). These inter-annual oscillations have a particularly strong effect on the strength and duration of seasonal precipitation and temperature conditions, as well as our ability to accurately forecast climate extremes in a spatially-explicit way. Once the CCYF is running using future climate scenario data, the next step would be to adjust such future scenario indicators based on a specific set of teleconnection strength scenarios.

5.5 Deploying in-season crop outlook reports with wireless mobile technology

Once the method has been; 1) validated to survey or other crop-insurance data to an agreed level, 2) future climate, teleconnection scenarios have been integrated, and 3) issues linked with use of remote-sensing data as auxiliary indices are resolved, then it would be very advantageous to release crop outlooks via a mobile application that users/producers can easily install remotely and receive updates from. This would incorporate the use of remote wireless technology in delivering this enhance decision-support directly to those who need it. At this stage, such technology could provide AAFC with key information on current and future level of anticipated use of the decision-support/forecasting tool, as well as encouraging further feedback and suggestions for enhancement linked with how crop outlooks are generated, visualized and explained. This would provide information needed to measure scientific 'impact' of the CCYF (current and future potential) i.e. number of users, when they are accessing crop outlooks through the growing season (user frequency graphs) as an indicator of when different producers need forecast information, user-ratings and guiding new ways to improve the understanding of crop outlook forecasts to a broad agricultural stake-holder audience. Other AAFC tools, such as the farm-scale whopper cropper decision-support tool, might also be released using for users via mobile wireless technology, leading to a suite of tools that AAFC supports and continues to enhance. This would represent a significant gain in bringing together the latest and greatest science alongside modern technology.

Acknowledgements

This research was funded under the Sustainable AGriculture Environmental Systems (SAGES) Program of Agriculture and Agri-Food Canada (AAFC), with additional support from AAFC's National Agro-climate Information Service (NAIS). We especially thank project collaborators: Allan Howard, Harvey Hill, Aston Chipanshi, Yinsuo Zhang, Andrew Davidson, Ian Jarvis, Heather McNairn and Lawrence Townley-Smith (Science and Technology Branch, AAFC) for their helpful suggestions, insights and guidance during the research design, coding and simulation testing of the integrated crop forecasting statistical method. We thank Budong Qian and Richard Warren (AAFC) for their assistance in acquiring and manipulating agro-climate input data and Andrew Davidson for processing and preparation of MODIS vegetation (NDVI) earth observational data. We also acknowledge input/feedback from a broader set of government collaborators involved in the operational development and delivery of the crop outlook prototype tool, as well as in assisting in validation work: C. Champagne, P. Cherneski, B. Daneshfar, A. Davidson, X. Geng, E. Gorelov, A. Howard, I. Jarvis, D. Qi, B. Qian, R. Rieger, L. Townley-Smith, D. Waldner and R.T. Warren From AAFC and F. Bedard and G. Reichert from Statistics Canada for their support and contributions.

References

Bastiaanssen, W.G.M., and Ali, S. (2003), "A new crop yield forecasting model based on satellite measurements applied across the Indus Basin, Pakistan", Agriculture, Ecosystems and Environment, 94, 321–340.

Bornn, L., and Zidek, J. V. (2012), "Efficient stabilization of crop yield prediction in the Canadian Prairies", Agricultural and Forest Meteorology, 152, 223-232.

Breiman, L. (2001), "Random Forests", Machine Learning, 45, 5–32.

Budong, Q., De Jong R., Warren, R., Chipanshi, A., Hill, H. "Statistical spring wheat yield forecasting for the Canadian Prairie provinces", Agricultural and Forest Meteorology, 149 (6-7), 1022-1031.

Challinor, A.J., Slingo, J.M., Wheeler, T.R., Craufurd, P.Q., Grimes, D.I.F. (2003), "Toward a combined seasonal weather and crop productivity forecasting system: determination of the working spatial scale", Journal of Applied Meteorology, 42, 175-192.

De Wit, A.J.W. and van Diepen, C.A., (2007), "Crop model data assimilation with the Ensemble Kalman filter for improving regional crop yield forecasts", Agricultural and Forest Meteorology, 146, 38-56.

ESRI, 2010. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute. URL: http://www.esri.com/

Hansen, J.W., Challinor, A., Ines, A., Wheeler, T., Moron, V., (2006), "Translating climate forecasts into agricultural terms: advances and challenges", Climate Research, 33: 27-41.

Hansen, J.W., Potgieter, A., Tippett, M.K., (2004), "Using a general circulation model to forecast regional wheat yields in northeast Australia", Agricultural and Forest Meteorology, 127, 77-92.

Hammer, G.L., Hansen, J.W., Phillips, J.G., Mjelde, J.W., Hill H., Love A., Potgieter, A. (2001), "Advances in application of climate prediction in agriculture". Agricultural Systems 70, 515-553.

Lansigan, F.P., Salvacion, A.R., Paningbatan, E.P. Jr., Solivas, E.S., and Matienzo, E.L.A. (2007), "Developing a Knowledge -based Crop Forecasting System in the Philippines", Proceedings of the 10th National Convention on Statistics (NCS), 14pp.

Joint Research Centre, European Commission. (2012), "MARS crop yield forecasting system (MCYFS)", http://www.marsop.info/marsop3/.

Khan, J., Aelst, S. V., Zamar, R. (2007), "Robust linear model selection based on least angle regression", Journal of the American Statistical Association, 102, 1289-1299.

Khan, J., Aelst, S. V., & Zamar, R. (2010). Fast robust estimation of prediction error based on resampling. Computational Statistics and Data Analysis, 54, 3121-3130.

Mkabela, M.S., Bullock, P., Raj, S., Wang, S., Yang, Y. (2011), "Crop yield forecasting on the Canadian Prairies using MODIS NDVI data", Agricultural and Forest Meteorology, 151(3), 385-393.

R Development Core Team. (2008). "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.

Rubas, D.J., Hill, H.S.J., Mjelde, J.W. (2006), "Economics and climate applications: exploring the frontier", Climate Research, 33, 43-54.

Semenov, M.A. and Doblas-Reyes, F.J. (2007), "Utility of dynamical seasonal forecasts in predicting crop yield", Climate Research, 34, 71-81.

Statistics Canada. (2007). "Census Agricultural Regions Boundary Files for the 2006 Census of Agriculture - Reference Guide", www.statcan.gc.ca/bsolc/olc-cel/olc-cel?lang=eng&catno=92-174-G.

Statistics Canada. (2012a). "1976-2011 Crops Small Area Data", Field Crop Reporting Series of Agriculture Division, Statistics Canada.

Statistics Canada. (2012b). "Definitions, data sources and methods of Field Crop Reporting Series, Record number: 3401", Agriculture Division, Statistics Canada., www.statcan.gc.ca/imdb-bmdi/3401-eng.htm

Stone, R.C., Meinke, H. (2005), "Operational seasonal forecasting of crop performance", Philosophical Transactions of the Royal Society B., 360, 2109-2124.

Wu, B. (2006), "Introduction of China CropWatch system with remote-sensing", ISRPS Archives XXXVI-8/W48.11, www.isprs.org/proceedings/XXXVI/8-W48/15_XXXVI-8-W48.pdf