Deep Learning for Improved Agricultural Risk Management

Nathaniel K. Newlands
Agriculture and Agri-Food Canada (AAFC)
nathaniel.newlands@canada.ca

Azar Ghahari
University of Texas at Dallas
azar.ghahari@utdallas.edu

Yulia R. Gel
University of Texas at Dallas
ygl@utdallas.edu

Vyacheslav Lyubchich
University of Maryland (UMCES)
lyubchich@umces.edu

Tahir Mahdi
Agriculture and Agri-Food Canada
tahir.mahdi@canada.ca

Abstract

Deep learning provides many benefits, including automation, speed, accuracy, and intelligence, and it is delivering competitive performance now across a wide range of real-world operational applications – from credit card fraud detection to recommender systems and customer segmentation. Its potential in actuarial sciences and agricultural insurance/risk management, however, remains largely untapped. In this pilot study, we investigate deep learning in predicting agricultural yield in time and space under weather/climate uncertainty. We evaluate the predictive power of deep learning, benchmarking its performance against more conventional approaches alongside both weather station and climate. Our findings reveal that deep learning offers the highest predictive accuracy, outperforming all the other approaches. We infer that it also has great potential to reduce underwriting inefficiencies and insurance coverage costs associated with using more imprecise yield-based metrics of real risk exposure. Future work aims to further evaluate its performance, from municipal area-yield, to finer-scale crop-specific producer-scale yield.

1. Introduction

1.1. Agricultural risk management

Risk analysis is a key feature for detecting geographical regions and time periods with the highest vulnerability of the target insurance market to specific natural hazards, and for identifying where the currently adopted risk management practices are ineffective and why – thus paving the way for the new insurance products in agriculture. The indemnity rate of an agricultural product (crop) moves up and down every year depending upon many risk factors such as weather, plant diseases, underwriting inefficiencies, technological changes and management practices. Losses due to underwriting inefficiencies are losses that can be decreased by making the underwriting process more efficient and through better management practice.

Recent work in applying deep learning to crop yield prediction consistently shows that this approach outperforms classical statistical approaches. This includes a recent study of county-level corn yield in the US Midwest with MSE error reductions of up to 32% [1]. You et al., (2017) also recently report on improved prediction skill with convolution neural networks (CNNs) achieving MSE error within 6%, compared to classical regression model benchmarks [2]. Mean MSE and MAE error estimates reported from the simulation of conventional crop growth/yield simulation models ranges between 2-30% [3, 4, 5].

1.2. Index-based insurance products

In principle, index products could be tailored to each insured client, however, in practice the data for constructing the index exhibit a limited spatial and temporal specificity. This is because the spatial resolution of the available agricultural and weather records is often inadequate for estimating a unique loss distribution for each insured client [6]. Index products can be classified into two main categories based on a type of index generalization, namely, *weather-based indices* and *aggregate loss indices* [7]. Payoffs for *weather indices* are based on a realization of one or several weather variables, e.g., rainfall, wind speed, and temperature, rather than the actual damage. *Aggregate loss indices* are based on expected losses over a group of individuals, e.g., an average area yield in the same geographical region. In both cases, the index represents a proxy for the losses of individual clients. A broad perspective of interacting agricultural and agri-food, health and development systems, considers both climatic and non-climatic risks, and multiple, complex pathways of interaction and response across supply chains [8, 9].

URI: https://hdl.handle.net/10125/59543
ISBN: 978-0-9981331-2-6

The following requirements are specified as key ones for an index product [7]: 1) an index should be highly correlated with the insured risk; 2) historical data archive for the index construction should be sufficiently long and represent a secure and objective data source, and 3) there should exist an indication of how the index is related to consequential losses and costs of the potential insurance clients. Since variability of the generalized index is lower than yield variability of an individual farm, spatial aggregation (over a group of clients) leads to errors in the risk assessment on a farm-level basis. This in turn leads to a poorer relation between the index and loss of insured assets (see the requirement 1 above), i.e., the effect known as the increased *basis risk*. While no product design can lead to a complete elimination of basis risk in index insurance, a rigorous statistical evaluation of sources of uncertainties and errors is critically important for reducing basis risk, minimizing losses to both insured and insurers, and for avoiding maladaptation. Levels of indemnity payments are determined from the realization of a particular index, such as accumulated temperature, precipitation, or based on expected losses over a group of individuals, rather than the actual loss. Advantages of index insurance include lower costs due to omitting the loss verification step, quicker claims settlement process, and elimination of fraud, adverse selection (i.e., hidden information), and moral hazard (i.e., hidden action).

1.3. Improving risk quantification

Weather reflects short-term conditions of the atmosphere while climate is the average daily weather for an extended period of time at a certain location[1]. Extreme weather that is both abrupt and challenging to predict, such as heat and drought, cause severe declines in crop yields and longer-term food insecurity. While drought is a prolonged, intensifying period of abnormal moisture deficiency; heat waves are episodic, and far more abrupt and difficult to predict in relation to lead time. Weather index-based crop insurance is being used by re-insurance companies like SwissRE and insurers like Climate Corporation (San Francisco, CA, USA).

A conditional approach (based on the Weibull distribution) has been applied to modeling yield risk (i.e., expected loss cost ratio) using producer-level corn yield data from Illinois, USA, to explore the impact of alternative weather and time horizons on yield risk estimation [10]. Their findings show that assumed risk evolution (i.e., increasing/decreasing) in rate-making methodologies can be violated leading to highly biased rates. Understanding the dynamic interaction between crop yields and various weather risks is, therefore, critically important in basis risk estimation and prediction [11]. Furthermore, Ghari et. al. (2017) have further explored variable selection techniques employing multi-resolution weather data in crop yield prediction for agricultural crop insurance [12]. The relationship between yields and weather variables is often nonlinear and intractable by a parametric description, exhibiting non-stationarity and non-separability of space-time covariances. Inference approaches that rely on conventional parametric models and assumptions, such as linear regression and spatial independence, may not accurately depict the underlying functional relationships. Significant efforts are being directed to developing new indices and index-based methodologies for assessing and modeling risk and use of indices as sustainability metrics [9, 13, 14, 15]. More recently, Dalhaus and Finger (2016) have assessed if gridded precipitation and crop phenological data can reduce basis risk of weather index–based insurance [16]. They found no differences between using gridded and weather station precipitation, but use of phenological data did increase expected utility. Also, Woodward (2016) has recently investigated the integration of high-resolution soil data into crop insurance policy design, reporting that the degree to which soils vary within a county is highly significant, leading to rating errors of 200% or greater [17].

1.4. Problem statement

Relatively few studies provide a comprehensive analysis of various uncertainty sources in agricultural insurance [14, 18, 19, 20]. Recent efforts to address these problems include the conditional distribution approach [10], copulas [18, 19, 21], wavelets, artificial neural networks, and other machine learning techniques (see the recent reviews [15, 22] and references therein), and Bayesian networks [23, 24]. These approaches are typically based on three types of data, such as losses of the insured, cause of loss, and the index at certain resolution scales. These scales are often too coarse for reliable risk quantification and impact assessment for decision makers, and geographical heterogeneity is often neglected. A new momentum in assessing risk and viability of index products is created by the availability of new multi-source, multi-resolution data, such as fine-grained remote sensing data, irregularly spaced observations at weather stations and gridded outputs from climate models on various scales [15]. Non-conventional modeling approaches, such as deep learning are well suited to these complex data

[1] https://oceanservice.noaa.gov/facts/weather_climate.html

types and are needed to integrate these multi-source disparate-resolution inputs into a reliable actuarial risk management framework.

1.5. Research objective

We present and discuss findings from a pilot study that compares the predictive performance of deep learning against other competing statistical approaches for index-based crop insurance. Our primary objectives are two-fold: to introduce deep learning methodology to agricultural index-based insurance, and to promote a greater dialogue about potential benefits (and challenges) of applying deep learning in agricultural crop insurance. We provide an overview of deep learning methodologies and inter-compare their predictive accuracy with other competing statistical approaches for Manitoba, Canada, a large agricultural region with a complex crop mix. We gauge the performance of deep learning methods in predicting crop yield based on two main types of climate data typically available for use by practitioners (i.e., disparate historical station-based observational versus interpolated and gridded climate reanalysis or reconstruction data). This paper is organized as follows. In Section 2, we describe the methods, their tuning parameters, and data we use. Section 3 provides the main results. Section 4 gives final comments and directions for the future work.

2. Methods and data sources

2.1. Problem definition

The core objective is to model crop yield (response variable Y) in relation to a set of climate and weather-related explanatory variables (predictors X): $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where, $\mathbf{Y}$ is the matrix of crop yields for i^{th} regions and t^{th} years ($i = 1, \ldots, n$; $t = 1, \ldots, T$), $\mathbf{X}$ is the design matrix comprising the X_{ijt} the j^{th} independent explanatory weather variables ($j = 1, \ldots, J$), and $\boldsymbol{\epsilon}$ is error (uncertainty) assumed to be white noise, i.e., $\boldsymbol{\epsilon} \sim \mathcal{WN}(0, \sigma^2)$. This equation is the simplest version of a generalized linear model (GLM) [25], where the expected value $E(\mathbf{Y}|\mathbf{X}) = \mu$, and $g(\mu) = \mathbf{X}\boldsymbol{\beta} = \mu$, i.e., the link function $g(x) = x$ is the identity function.

2.2. Benchmark models

The following benchmark models were selected as competitors for deep learning: GLM and its two modifications based on the model selection algorithms of screening regression (SR) and principal component analysis screening regression (PCASR) [11, 12], gradient boosting (GB) [26] and random forest (RF) learning [27].

The SR and PCASR models employ dimension reduction to address multicollinearity in explanatory variables and singularity of the design matrix. SR iteratively selects (screens) predictors having the highest correlation with crop yield. PCASR performs a PCA transformation of the $\mathbf{X}_i$ design matrix before applying the screening to its principal components.

RFs use "bagging" – bootstrap aggregation – of regression trees to decrease variance and so are suitable for high variance and low bias problems, whereas GB employs "boosting" or weighted ensemble averaging of trees to decrease bias, and is suitable for low variance and high bias problems.

Deep learning employs "stacking" (as opposed to bagging or boosting). This approach comprises a set of machine-learning algorithms that use a neural network architecture, involve nonlinear functions, and have no set requirement of number of layers or variables. We use neural networks with many hidden layers (deep neural networks or DNNs) and graphical or belief models (deep belief networks or DBNs) formed by many levels of hidden variables. Feedforward neural networks (FNNs) have the simplest network structure, with directed linkages between the input, hidden and output layers, and no linkages between nodes in each layer. The FNN model relates activations a_j^l of the j^{th} neuron in the l^{th} layer to the summation of all activations over all neurons, k, in the $(l-1)^{th}$ layer, given by:

$$a_j^l = f\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right), \tag{1}$$

where, w_{jk}^l is the *weight* of the connection (or link) from the k^{th} neuron (or node or unit) in $(l-1)^{th}$ layer to the j^{th} neuron in the l^{th} layer, b_j^l is the *bias* for the j^{th} neuron in the l^{th} layer, and a_j^l is the *activation* of the j^{th} neuron in the l^{th} layer. Furthermore, $f(\cdot)$ is termed the *activation function* and w^l is termed the network *weight* matrix, i.e., weights of the l^{th} layer.

Given a general objective, cost or energy function, E, the error vector for a given layer l is the derivative with respect to its weighted input, $\epsilon_j^l = \partial E / \partial z_j^l$, and supervised learning (i.e., via backpropagation that requires the activation function to be differentiable) computes this error backwards starting from the final layer in relation to the gradients of the energy function with respect to neural weights and biases, namely,

$\partial E/\partial w_{jk}^{l}$ and $\partial E/\partial b_{j}^{l}$. By applying the chain rule, it can be shown that the weight update rule via gradient descent to minimize network energy E and to obtain a more "desirable" model configuration, is then, $\Delta w_{ij} = -\eta \partial E/\partial w_{ij}$, where η is the learning rate. A learning rate that is too high can cause the network optimization to miss the global optimum, and one that is too low, can slow convergence. Most of the recent experimental results with a large number of hidden layers are obtained with models that can be turned into deep *supervised* neural networks, but with initialization or training schemes that differ from FNNs [28].

New algorithms that use *unsupervised* pre-training have been found to perform better than those that use standard random initialization and gradient-based optimization. Unsupervised pre-training acts as a regularizer that initializes the parameters in a "better" basin of attraction of the optimization procedure, corresponding to an apparent local minimum associated with better generalization. This form of model training is a form of regularization that minimizes variance, while introducing bias. It modifies the initialization parameters for supervised training, rather than modifying the objective function [29].

Let visible and hidden units be denoted v and h, respectively, and $\theta = (bw, v, h)$, for weight matrix w. In training a single restricted Boltzmann machine (RBM), weight updates are performed using gradient descent with alternating Gibbs sampling, whereby,

$$\Delta\Delta w_{ij} = -\eta \frac{\partial \ln(p(v,\theta))}{\partial w_{ij}}, \tag{2}$$

where $p(v)$ is the probability of a visible vector. The joint probability of the visible, hidden layers, bias and weight model parameters is given by:

$$\begin{aligned} p(v,h,\theta) &= \frac{1}{Z(\theta)} \exp\big(-E(v,h,\theta)\big) \\ &= w_{ij}(t) + \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model} \end{aligned} \tag{3}$$

for a partition function Z and energy function, $E(v,h)$ of a given network state. A pre-training step maximizes the log-likelihood of the data, denoted $\ln(p(v,\theta))$, using the contrasting divergence (CD) approach. The expected value of distribution p is denoted by $\langle \cdot \rangle_p$. The CD approach approximates the log-likelihood gradient and has been found to be a successful weight update rule for training RBMs [30, 31]. We then proceed as follows:

1. Initialize visible units to a training vector.
2. Update hidden units in parallel given the visible units.
3. Apply a reconstruction step that updates the visible units given the hidden units.
4. Re-update the hidden units given the reconstructed visible units.
5. Apply the weight update $\Delta w_{ij} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}$.

An optimal DBN network structure is found based on learning scheme and parameters, number of hidden layers, activation function type, momentum, etc. (for definition of these different parameters, we refer readers to the practical guide by [32]).

2.3. Agricultural yield data

The province of Manitoba was selected as the study region based on availability of crop yield and climate information and for comparison with the results of [11]. Historical, farm-scale panel crop yield data (yield per acre and total acres, 1996–2011, where the units of yield are metric tonnes, mt) covering 99 regional municipalities, 75 crop types from 19,238 farms were used, publicly available online from the Manitoba Agricultural Services Corporation (MASC) MMPP Variety Yield Data Browser[2]. Manitoba municipalities vary substantially by their harvested area, from 2 to 3,572 km^2.

A consistent set of crop yield data was extracted for 53 crops for which yield data was available across all the municipalities. The yield for each crop in a given municipality was then re-scaled within the interval $[0, 1]$ based on its maximal historical yield (1996–2011). Such normalization has two linked benefits: 1) comparatively assesses (risks of) low yields for different crops, and, 2) enables aggregation across different crops. A re-statement procedure was next applied to the re-scaled yields using the same procedure detailed in [11]. This aggregates yield estimates for each single crop as a weighted-average representative crop mix, so as to provide a better reflection of the current risk profile as crop selection differs from year to year. This corrected historical yield data for uncertainty introduced by changes in farming practices such as changes in crop mix, crop rotation, and mixed cropping, helping to ensure that the historical observations are sufficiently representative of current production as well as good indicators of future crop production. The response variable, Y, is thus a weighted average yield for the representative crop mix (i.e., rather than for a specific crop type or variety) for each municipality and covering at least 90% of the area in the most recent five years.

[2] https://www.masc.mb.ca/mmpp2.nsf/mmpp_browser_variety.html

2.4. Climate and weather data

Historical, daily climate data (minimum, maximum, average temperature and total precipitation) for 38 high-quality stations situated on agricultural land across Manitoba were obtained from Environment Canada's archive of the Adjusted and Homogenized Canadian Climate Data (AHCCD).

These observational station-based data, were complemented by gridded climate reanalysis data obtained from the ERA-Interim archive that provided higher coverage of the study region (see Fig. 1 in [12]). ERA-Interim is a global atmospheric reanalysis from 1979, continuously updated in real time produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) [33]. The ERA-Interim data assimilation system uses 4-dimensional variational analysis with a 12-hour window, and spatial horizontal resolution of approximately 80 km, with 60 vertical levels from the surface up to 0.1 hPa.

Gridded data was used daily average temperature at 2 m and total precipitation from the ERA-Interim archive for a subset of 16 years (1996–2011) that matched the available historical crop yield data. Centroids of the Manitoba regional municipality polygon areas were used to obtain values of the variables from the ERA reanalysis grid. Referencing centroid values could potentially introduce error, since the gridded data are calculated at each vertex of the computational mesh; while the data have relatively low spatial frequency, this could potentially introduce a spatial phase error up to 40 km.

Following [11], two matrices of weather indices – separately based on the observed station and reanalysis datasets – were computed for all 16 years (1996–2011) and municipalities, i.e., annual indices, for each month in the growing season (May–October). Each of the matrices included up to 432 variables: night- and day-growing degree low (NGDL, DGDL), night- and day-growing degree high (NGDH, DGDH), precipitation low and high (PREH, PREL), etc. The matrices were cleaned by removing constant indices (i.e., those with only one unique value per index) and duplicate indices, if any. The resulting quality-controlled weather-index matrix based on observed station data contains 333 variables, and based on reanalysis data – 292 variables. Whereas the reanalysis data provide better spatial coverage than weather stations, we use only the municipalities for which station-based weather data are available, to ensure comparability of the results.

2.5. Model implementation

2.5.1. Algorithms The R package *h2o* was used for this analysis [34]. This package provides fast scalable open source tools for machine learning and deep learning, using in-memory compression techniques, Hadoop and Spark cluster-based computing, having a platform that interfaces with R, Python, Scala, Java, and more. It implements GLMs, naive Bayes, PCA, time series analysis, k-means clustering, RFs, GB, and deep learning [35]. *h2o* follows the model of multi-layer, feed-forward neural networks for predictive modeling and uses an improved stochastic gradient descent. The DBN model was implemented using the R *deepnet* library [36].

2.5.2. Regularization Fitting a training set too closely can limit the ability to generalize a model - termed overfitting. It can be reduced by constraining the fitting procedure using so-called regularization techniques, such as: ridge regression and LASSO (least absolute shrinkage and selection operator) [37], elastic net penalty, stochastic training of gradient boosting [25, 26, 38], penalized complexity, and shrinkage with a fixed or adaptive learning rate. In the case of deep learning, "dropout" is the technique used to prevent overfitting and to combine exponentially many different neural network architectures in an efficient way to find optimal hyperparameters for many different model architectures. Dropout randomly selects and temporarily removes nodes (visible or hidden nodes, along with their connections) from a network during training, creating an exponential number of different "thinned" networks that are then sampled as part of an approximate averaging method. Dropout provides a significantly lower generalization (out-of-sample) error, compared to other regularization methods that can be used in training, and improves the performance of neural networks for supervised learning on many different benchmark datasets [39].

2.5.3. Validation To assess predictive performance of the models, we use leave-one-year-out (LOYout) cross-validation:

1. For year $t' = 1, \ldots, T$:
   (a) Use Y_{it} and X_{ijt} ($t \neq t'$) to train the models.
   (b) Use $X_{ijt'}$ to forecast yields for all municipalities ($i = 1, \ldots, n$) for the year t': $\widehat{Y}_{it'}$.
   (c) Compute absolute errors for year t': $\mathrm{AE}_{it'} = \left| Y_{it'} - \widehat{Y}_{it'} \right|$.
   (d) Compute squared errors for year t': $\mathrm{SE}_{it'} =$

$$\left(Y_{it'} - \widehat{Y}_{it'}\right)^2.$$

2. Mean errors for each municipality:

$$\mathrm{MAE}_i = T^{-1}\sum_{t=1}^{T}\mathrm{AE}_{it};\ \mathrm{MSE}_i = T^{-1}\sum_{t=1}^{T}\mathrm{SE}_{it}. \tag{4}$$

3. Overall mean errors:

$$\mathrm{MAE} = n^{-1}\sum_{i=1}^{n}\mathrm{MAE}_i;\ \mathrm{MSE} = n^{-1}\sum_{i=1}^{n}\mathrm{MSE}_i. \tag{5}$$

MAE and MSE are yield-based error metrics comparing actual versus model predicted yield, alongside other metrics used in agricultural risk management/crop insurance such as the Loss-Cost (LCR) that uses an estimate of average or expected yield and to compute premium pricing rates for various coverage ratios (c) and risk loading levels (θ).

2.5.4. Sensitivity and Robustness Typical model hyperparameters include: choice of activation function (e.g., tanh, rectifier, or maxout), size and number of hidden layers, number epochs (i.e., number of iterations for model tuning), number of folds for cross-validation, dropout ratio (to improve generalization error), stopping tolerance, learning rate, and momentum.

The GLMs assumed normal or Gaussian-distributed variables. The SR considered 30% of variables with the highest absolute value of correlation coefficient with the response. The PCASR model considered principal components explaining at least 85% of the variation. The GB comprised a total of 50 trees, a learning rate of 0.1, and stopping tolerance of 0.001. The RF had a total of 10 trees, with a maximum tree depth of 4, and default stopping tolerance of 0.001 for residual deviance. The DNNs considered configurations having three hidden layers of 5 units each, and two hidden layers with 10 units each, with total epochs varying from 10 to 50 (i.e., number of times to iterate a dataset), a total of 10,000 validation set samples for scoring, and a score duty cycle of 0.005 (to promote training), with a fixed learning rates of 0.01 and 0.05 with annealing rate of $2 \cdot 10^{-6}$. The DBN model used a tanh-type activation function, with two hidden layers of 5 units, with 5 epochs, batch size of 100, and a learning rate of 0.5. While reasonable default values of model hyperparameters are specified, tuning maximizes predictive model performance. *h20* provides manual or automated re-tuning of model hyperparameters in its cross-validation procedure. In the case of the FFN and DNN models implemented using *h20*, a grid search (i.e., Cartesian hyperparameter search or exhaustive search) algorithm is invoked that builds a separate model under every combination of hyperparameter values set manually by a user or determined automatically based on a given validation metric (e.g., MSE or MAE). Multiple searches can be run with results on all the runs collected for comparison in a single set, keeping track of all these models resulting from the search. For the DBN model, a manual hyperparameter search was conducted by varying the number of hidden layers (2–10) and learning rate.

For each model, separate scenario runs were performed using the different input datasets to check model robustness to varying spatial heterogeneity and the different types of training climate and weather data: 1) observed station-based data, 2) ERA-Interim reanalysis data, and 3) combination of the two datasets.

3. Results

LOYout cross-validation (Section 2.5.3) was applied to the seven competing models. Cross-validated error statistics (5) for each model are summarized in Table 1. Deep learning (DNN and DBN) out-performed all other models, with DNN having the lowest prediction errors. In turn, the GLM had the highest prediction errors for the weather station dataset and combined datasets, and highest MAE for the reanalysis dataset. For the reanalysis dataset, PCASR model had a higher MSE than the GLM. The tree-based methods had a greater accuracy than the GLM, SR, and PCASR models in most of the cases. Bagging (RF) performed better than boosting (GB) in terms of MAE for the case of station and reanalysis data, but not for the combined data.

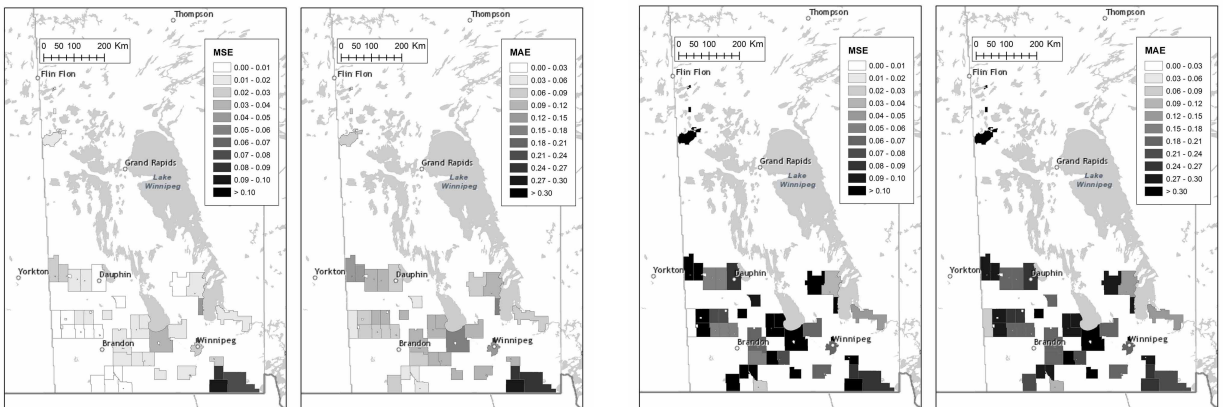
The yields in municipalities within the south-eastern portion of the study region prove difficult for all models to predict from the weather indices, while the GLM shows the largest error spread. Figures 1 and 2 show spatial maps of model prediction bias (MAE) and error variance (MSE) for the best-performing deep learning DNN model compared to the GLM benchmark. Results obtained using weather station and climate reanalysis data are provided. Overall, the observed weather data capture observed yield variability better than the climate reanalysis data.

Tables 2 and 3 show results for the same type of methods, but applied to only two municipalities (Ellice and Hamiota) with sparse station-based weather data. In the situation of scarce data, some of the baseline methods may fail completely (e.g., see the performance of SR in Tables 2 and 3), however, the deep learning approaches show consistently better performance than other considered methods. Furthermore, by comparing different columns within each Table 2 and 3, we observe that GLM, SR, and PCASR have

Table 1. Error in model predictions (i.e., cross-validated model bias, MAE, and model error variance, MSE) for different weather/climate information/datasets. The deep learning models (i.e., DNN and DBN models) show the best performance with lowest error.

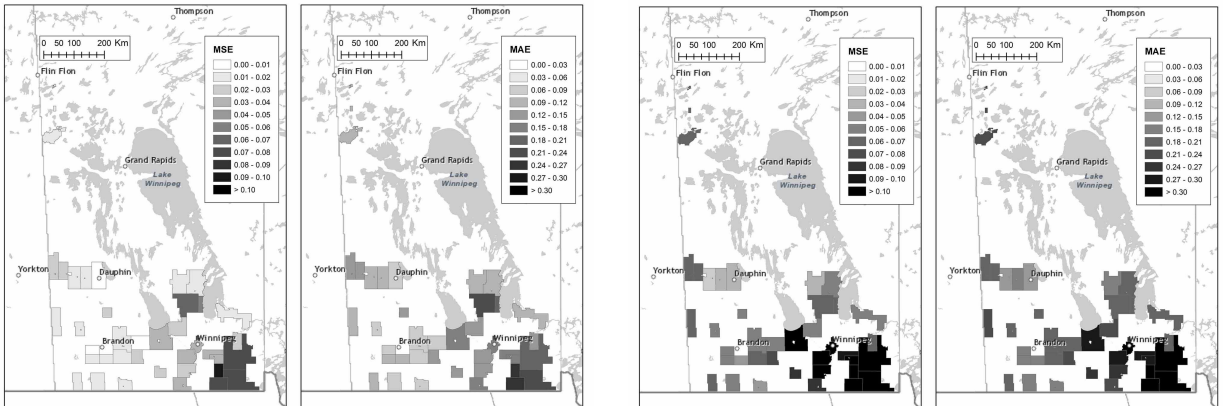
	Station-Based Weather Data		Climate Reanalysis Data		Combined Weather Data	
Model	MSE	MAE	MSE	MAE	MSE	MAE
GLM	0.0539	0.1733	0.0732	0.2115	0.0402	0.1512
SR	0.0367	0.1506	0.0428	0.1580	0.0312	0.1381
PCASR	0.0277	0.1267	0.1125	0.1864	0.0281	0.1287
GB	0.0259	0.1258	0.0293	0.1316	0.0270	0.1239
RF	0.0274	0.1252	0.0307	0.1333	0.0286	0.1292
DNN	0.0248	0.1202	0.0263	0.1230	0.0262	0.1213
DBN	0.0253	0.1228	0.0268	0.1231	0.0264	0.1255

Note: Methods are: generalized linear model (GLM), screening regression (SR), principal component analysis screening regression (PCASR), gradient boosting (GB), random forest (RF), deep neural network (DNN), and deep belief network (DBN). The mean squared and mean absolute errors (MSE and MAE) are calculated using formulas (5), number of municipalities n is 34; number of years T is 16.



(a) Deep Neural Network (DNN) (b) Generalized linear model (GLM)

Figure 1. Cross-validated errors (MSE, MAE) for the best-performing DNN model versus the GLM benchmark for *weather station-observed data*.



(a) Deep Neural Network (DNN) (b) Generalized linear model (GLM)

Figure 2. Cross-validated errors (MSE, MAE) for the best-performing DNN model versus the GLM benchmark for *climate reanalysis data*.

Table 2. Cross-validated error results summary for the Ellice municipality

	Station-Based Weather Data		Climate Reanalysis Data		Combined Weather Data	
Method	MSE	MAE	MSE	MAE	MSE	MAE
GLM	0.0087	0.0752	0.0278	0.1388	0.4559	0.2163
SR	0.0712	0.1392	30.9591	2.8387	473.9329	6.9779
PCASR	0.0089	0.0771	0.3348	0.2266	0.3689	0.1842
GB	0.0089	0.0792	0.0096	0.0757	0.0099	0.0791
RF	0.0103	0.0858	0.0138	0.1008	0.0102	0.0789
DNN	0.0070	0.0695	0.0058	0.0623	0.0073	0.0726
DBN	0.0071	0.0711	0.0067	0.0666	0.0098	0.0871

Note: The mean squared and mean absolute errors (MSE and MAE) are calculated using formulas (4), number of years T is 16.

Table 3. Cross-validated error results summary for the Hamiota Municipality

	Station-Based Weather Data		Climate Reanalysis Data		Combined Weather Data	
Method	MSE	MAE	MSE	MAE	MSE	MAE
GLM	0.0162	0.0953	0.0332	0.1298	0.3826	0.2324
SR	8.1738	1.9411	6.7312	1.6779	20923.2019	51.7363
PCASR	0.0115	0.0845	0.2725	0.2288	0.2973	0.1757
GB	0.0111	0.0822	0.0164	0.0972	0.0146	0.0914
RF	0.0165	0.0986	0.0153	0.0900	0.0131	0.0852
DNN	0.0101	0.0753	0.0111	0.0836	0.0113	0.0780
DBN	0.0107	0.0757	0.0115	0.0738	0.0130	0.0794

Note: The mean squared and mean absolute errors (MSE and MAE) are calculated using formulas (4), number of years T is 16.

unstable performance, when the climate reanalysis or multi-resolution (combined) data are used, instead of the station-based data. Using the combined data does not lead to lower generalization errors in these two cases. The stable performance of the deep learning methods, in particular, DNN and DBN, suggests a better robustness of these approaches when used with a variety of data inputs.

4. Conclusions

The effectiveness of deep learning (and other machine-learning) algorithms has been discovered in a variety of real-world applications, such as recent work using DNNs, GB, RFs, and various ensembles of these methods in the context of statistical arbitrage and forecasting in economic markets [40]. Statistical arbitrage refers to quantitative trading strategies generally deployed within hedge funds or proprietary trading desks that are considered to be systematic (rule-based), market-neutral trade, and excess returns are statistically generated. Kraus et al. (2017) show that RFs outperform GB models and DNNs, but indicate that further hyperparameter optimization may still yield advantageous results for the tuning-intensive DNNs [40]. In contrast, DNNs outperformed all the other approaches in terms of predictive power in our study, with the DBN approach showing very similar results.

Crop models that are highly complex, with hundreds of parameters and variables, and assume interactions among system components that are largely empirical. These simulation models also have substantial structural inconsistency which is attributed to be model design and calibration error, Furthermore, there does not seem to be any simple relationship between model structure or the approach used to simulate individual processes and model error [4]. The results obtained here for DNN and DBN show accuracy gains are possible with reduced model complexity and computational efficiency conferred by deep learning. MSE and MAE average error estimates obtained from the simulation of these conventional crop models vary widely between (2-30)%. The validation of four widely-applied crop yield simulation models, under varying climate, cultivar, and sowing date effects, namely, CROPSIM-Wheat, CERES-Wheat, Nwheat, and APSIM-Wheat show an average MSE of 47.81% and average MAE of

1.59% [3]. Also, average MSE in grain yield (GY) from the validation of 27 crop (wheat) models is 9% (Figure 4a, pg. 919 of [4]). Average MSE error (4.91-5.60)% (or MSE of 0.00241-0.00314) and MAE error (18.12-19.42)% or (0.01812-0.1942) has been estimated from the validation of multiple linear regression, and machine-learning methods (i.e., M5-Prime regression trees, perceptron multilayer neural networks, support vector regression and k-nearest neighbor methods) for crop yield prediction against 10 crop datasets [5].

Our findings indicate that deep learning has the potential to reduce crop insurance underwriting inefficiencies and coverage costs associated with using more imprecise yield-based metrics of real risk exposure. Deep learning (supervised, DNN and unsupervised, DBN) both outperformed the benchmark models when applied to the annual crop yield and daily weather multi-scale data. The deep learning results (i.e. for the DNN and DBN models) range between (0.2-0.5)% for MSE and (0.3-.0.4)% for MAE across all the three different data input cases. These values are close. However, comparing the deep learning model results to the classical statistical GLM benchmark, for example, yields an error deviation of roughly (1-5)% for MSE, and (3-9)% for MAE, across the three input data cases, or a roughly 3-fold increase in mean MAE and MSE error. Further testing of the benchmark and deep learning models against climate and crop yield test data from other regions over longer time periods is, however, needed to more reliably quantify the expected performance benefits of deep learning in crop yield prediction. The findings reported here, while limited to the test data, show that deep learning models can outperform a suite of traditional benchmark models, and such methods strongly warrant further, broader attention and consideration for crop yield prediction.

Our findings are supported by those of the recent study of county-level corn yield in the US Midwest showing deep learning outperforms classical methods reducing MSE up to 32% [1]. Our findings are also consistent with the findings of You et al. (2017) who report on improved prediction skill with convolution neural networks (CNN's), achieving MSE error within 6%, compared to classical regression model benchmarks [2]. This predictive skill estimate for CNN's is within the range we have found with DNN and DBN, but a wider difference may arise in other applications. Other reanalysis datasets such as North American Regional Reanalysis (NAAR) could also be utilized. The current evaluation relies on mean estimates of the MSE and MAE error, instead of bootstrapped predictive confidence intervals. The validation metrics (MSE, MAE) are also point-wise, whereas a more rigorous approach would involve error modeling to evaluate spatial bias, dependence, nonlinearity between the municipalities in relation to topography, soil type, and other agricultural landscape variables [41]. Other model performance metrics, such as reduced complexity and computational efficiency, could be considered.

Our pilot study focused on weather and climate variables, but additional environmental covariates likely would help to improve prediction skill. Considering other variables like soil water content could change our prediction skill results. Soil data from reanalysis data could also be considered, which would also require assimilating soil probe observational data. Future work aims to further evaluate the performance of deep learning at the regional (i.e., municipal) scale and finer, producer-scale using additional reanalysis datasets and a CNN modeling approach.

5. Acknowledgements

YRG and VL were funded by the Society of Actuaries (SOA); NKN and TM by AAFC.

References

[1] A. Crane-Droesch, "Machine learning methods for crop yield prediction and climate change impact assessment in agriculture," *Environmental Research Letters*, forthcoming.

[2] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep gaussian process for crop yield prediction based on remote sensing data," Proceedings of the 31th AAAI Conference on Artificial Intelligence (AAAI-2017), 4-9 February, San Francisco, CA, USA, 2017.

[3] J. Hussain, T. Khaliq, A. Ahmad, and J. Akhtar, "Performance of four crop model for simulations of wheat phenology, leaf growth, biomass and yield across planting dates.," *PLoS ONE*, vol. 13, no. 6, p. e0197546, 2018.

[4] P. Martre *et al.*, "Multimodel ensembles of wheat growth: many models are better than one.," *Global Change Biology*, vol. 21, p. 911925, 2015.

[5] A. Gonzalez-Sanchez, J. Frausto-Solis, and W. Ojeda-Bustamante, "Predictive ability of machine learning methods for massive crop yield prediction.," *Spanish Journal of Agricultural Research*, vol. 21, no. 2, pp. 313–328, 2014.

[6] H. Ibarra and J. Skees, "Innovation in risk transfer for natural hazards impacting agriculture," *Environmental Hazards*, vol. 7, pp. 62–69, 2007.

[7] J. Skees, "State of knowledge report – data requirements for the design of weather index insurance," tech. rep., GlobalAgRisk, Inc., 2010.

[8] A. J. Challinor, W. N. Adger, and T. G. Benton, "Climate risks across borders and scales," *Nature Climate Change*, vol. 7, pp. 621–623, 2017.

[9] N. K. Newlands, *Future Sustainable Ecosystems: Complexity, Risk, Uncertainty*. Applied Environmental

Statistics, Baton Rouge, FL: Chapman & Hall/CRC, 2016.

[10] J. D. Woodward, "Impacts of weather and time horizon selection on crop insurance ratemaking: A conditional distribution approach," *North American Actuarial Journal*, vol. 18, no. 2, pp. 279–293, 2014.

[11] W. Zhu, L. Porth, and K. Tan, "A credibility-based yield forecasting model for crop reinsurance pricing and weather risk management," *SSRN(6)*, pp. 1–36, 2015.

[12] A. Ghahari, Y. R. Gel, V. Lyubchich, Y. Chun, and D. Uribe, "On employing multi-resolution weather data in crop insurance," in *Proceedings of the SIAM International Conference on Data Mining (SDM17) Workshop on Mining Big Data in Climate and Environment*, (Houston, Texas, USA), SIAM, April 2017.

[13] J. Skees, A. Murphy, B. Collier, M. J. McCord, and J. Roth, "Scaling up index insurance," tech. rep., GlobalAgRisk, Inc. & Microinsurance Centre, LLC, 2007.

[14] J. de Leeuw, A. Vrieling, A. Shee, C. Atzberger, K. M. Hadgu, C. M. Biradar, H. Keah, and C. Turvey, "The potential and uptake of remote sensing in insurance: A review," *Remote Sensing*, vol. 6, no. 11, pp. 10888–10912, 2014.

[15] L. Porth and R. Villeneuve, "Issues in agricultural insurance," in *Proceedings of the 2015 SOA Annual Meeting and Exhibit*, (Austin, TX), SOA, October 2015.

[16] T. Dalhaus and R. Finger, "Can gridded precipitation data and phenological observations reduce basis risk of weather indexbased insurance," *American Meteorological Society*, vol. 8, pp. 409–419, 2016.

[17] J. Woodward, "Integrating high resolution soil data into federal crop insurance policy: Implications for policy and conservation," *Environmental Science & Policy*, vol. 66, pp. 93–100, 2016.

[18] D. Nadolnyak and D. Vedenov, "Information value of climate forecasts for rainfall index insurance for pasture, rangeland, and forage in the Southeast United States," *Journal of Agricultural & Applied Economics*, vol. 45, pp. 109–124, 2013.

[19] B. K. Goodwin and A. Hungerford, "Copula-based models of systemic risk in u.s. agriculture: Implications for crop insurance and reinsurance contracts," *American Journal of Agricultural Economics*, vol. 97, no. 3, pp. 879–896, 2014.

[20] J. Martinich, A. Crimmins, R. H. Beach, A. Thomson, and J. McFarland, "Focus on agriculture and forestry benefits of reducing climate change impacts," *Environmental Research Letters*, vol. 12, no. 6, p. 060301, 2017.

[21] M. J. Miranda and K. Farrin, "Index insurance for developing countries," *Applied Economic Perspectives and Policy*, vol. 34, no. 3, pp. 391–427, 2012.

[22] L. Porth and K. S. Tan, "Agricultural insurance – more room to grow?," *The Actuary Magazine*, vol. 12, no. 2, pp. 35–41, 2015.

[23] N. K. Newlands and L. Townley-Smith, "Predicting energy crop yield using Bayesian networks," in *Proceedings of the Fifth IASTED International Conference*, vol. 711, pp. 14–106, 2010.

[24] J. D. Daron and D. A. Stainforth, "Assessing pricing assumptions for weather index insurance in a changing climate," *Climate Risk Management*, vol. 1, pp. 76–91, 2014.

[25] J. J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, no. 1, 2009.

[26] T. Hastie, R. Tibshirani, and J. J. H. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2001.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, vol. 9, (Sardinia, Italy), pp. 249–256, Proceedings of Machine Learning Research (PMLR), 2010.

[29] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?," *Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.

[30] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[31] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," Tech. Rep. GCNU TR 2000-004, Department of Computer Science, University of Toronto, 2002.

[32] G. E. Hinton, "A practical guide to training restricted boltzmann machines," Tech. Rep. UTML TR 2010-003, Department of Computer Science, University of Toronto, 2010.

[33] D. P. Dee, "The ERA-interim reanalysis: Configuration and performance of the data assimilation system," *Quarterly Journal of the Royal Meterological Society*, vol. 137, no. 656, pp. 553–597, 2011.

[34] The H2O.ai team, *h2o: R Interface for H2O*, 2017. R package version 3.10.5.3. `https://CRAN.R-project.org/package=h2o`.

[35] A. Candel, J. Lanford, E. LeDell, V. Parmer, and A. Arora, "Deep learning with h20. 3rd ed.," 2015. `http://h2o.gitbooks.io/deep-learning/`.

[36] X. Rong, *Deep Learning Toolkit in R*, 2015. Version 0.2. `https://CRAN.R-project.org/package=deepnet`.

[37] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[38] J. J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[39] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[40] C. Krauss, X. A. Do, and N. Huck, "Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the s&p 500," *European Journal of Operational Research*, vol. 259, pp. 689–702, 2017.

[41] Y. Tian, G. Nearing, C. Peters-Lidard, K. Harrison, and L. Tang, "Performance metrics, error modeling, and uncertainty quantification," *American Meteorological Society*, vol. 144, no. 6, pp. 607–613, 2016.