Learning Space-Time Crop Yield Patterns with Zigzag Persistence-Based LSTM: Toward More Reliable Digital Agriculture Insurance

Tian Jiang[1], **Meichen Huang**[1], **Ignacio Segovia-Dominguez**[1, 3], **Nathaniel Newlands**[2*], **Yulia Gel**[1, 4]

[1]Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080, USA
[2] Summerland Research and Development Centre, Science and Technology Branch, Agriculture and Agri-Food Canada, Summerland, BC, V0H1 1Z0, Canada
[3] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA
[4] Energy Storage and Distributed Resources Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
Tian.Jiang@utdallas.edu, Meichen.Huang@UTDallas.edu, Ignacio.SegoviaDominguez@UTDallas.edu, nathaniel.newlands@agr.gc.ca, ygl@utdallas.edu

Abstract

More than US$ 27 billion is estimated to have been paid-out in farm support in USA alone since 1991 in response to climate change impacts on agriculture, with costs likely continuing to rise. With the wider adoption of precision agriculture – an agriculture management strategy that involves gathering, processing and analyzing temporal, spatial and individual data – in both developed and developing countries, there is an increasing opportunity to harness accumulating, shareable, big data using artificial intelligence (AI) methods, collected from weather stations, field sensor networks, Internet-of-Things devices, unmanned aerial vehicles, and Earth observational satellites. This requires smart algorithms tailored to agricultural data types, integrated into digital solutions that are viable, flexible, and scalable for wide deployment for a wide variety of agricultural users and decision-makers. We discuss a novel AI approach that addresses the real-world problem of developing a viable solution for reliably, timely, and cost-effectively forecasting crop status across large agricultural regions using Earth observational information in near-real-time. Our approach is based on extracting time-conditioned topological features which characterize complex spatio-temporal dependencies between crop production regions and integrating such topological signatures into Long Short Term Memory (LSTM). We discuss utility and limitations of the resulting zigzag persistence-based LSTM (`ZZTop-LSTM`) as a new tool for developing more informed crop insurance rate-making and accurate tracking of changing risk exposures and vulnerabilities within insurance risk areas.

Introduction

Accurate crop monitoring and forecasting is crucial for ensuring food security and sustainable development. Agricultural landscapes are highly exposed to extreme weather events attributed to climate change that are becoming more frequent and intense such as flooding, heatwaves, and prolonged drought. Such events, coupled with underlying climate change trends, disrupt socio-environmental systems, which alters nutrient and water availability, invasive and beneficial pest populations, and soil microbe biodiversity. Crop growth is influenced by genotypes, weather dynamics, soil properties as well as agronomic management, that is, a wide variety of interdependent factors whose sophisticated spatio-temporal dynamics often cannot be jointly addressed with more traditional methods. In turn, artificial intelligence (AI) approaches, including deep learning (DL) tools, have a potential to more accurately capture such socio-environmental dependence patterns. The AI implementations based on DL are increasingly widely applied in agriculture alongside the broader use of diverse data from mobile and fixed field sensor networks, smartphone and internet-of-things (IoT) devices for planting, monitoring, assessing, protecting, and harvesting crops and livestock. For instance, such recent studies include DL for cropland classification (Jia et al. 2019), crop growth stage estimation (Worrall, Rangarajan, and Judge 2021), analysis of yield expectations (Shoshi et al. 2021), and agricultural commodity prices (Guo, Woodruff, and Yadav 2020). van Klompenburg, Kassahun, and Catal (2020) provide a detailed review of machine learning (ML) in crop yield prediction.

Furthermore, the emerging concept of *smart farming* utilizes AI to analyze information collected with sensors, drones and satellites to improve agricultural production and management. This automation system must consider variation in both environmental conditions and crop features to make timely and robust decisions. Meteorological information is often summarized in indices such as growing degree days (GDD) and heating degree days (HDD), which show strong predictive power in crop yield studies (Jiang et al. 2019; Zhu, Porth, and Tan 2019). There are also weather index derivatives traded in some exchanges as well as over-the-counter markets. The Chicago Mercantile Exchange (CME) provides standardized weather futures contracts since 1999 and now extends to nine US cities, two European cities, and one Japanese city based on weather indexes such as heating degree days (HDDs), cooling degree days (CDDs) and cumulative average temperature (CAT)[1]. With these financial instruments, insurance companies can potentially hedge climate change and extreme weather risks, which facilitates further development of index-based agricultural insurance products. Indeed, crop yield data collected

*Correspondence Author

[1]CME Group https://www.cmegroup.com/trading/weather/

through, and at the end of a growing season, traditionally has been collected using farm surveys that are expensive, labour intensive, yet prone to substantial measurement error, including overestimation bias. The availability of satellite imagery is cost efficient, of high spatial and temporal resolution, and standardized with higher objectivity being less dependent on reporting error and biases. Widely applied satellite-based remote sensing indices include normalized difference vegetation index (NDVI) (Quarmby et al. 1993), green chlorophyll vegetation index (GCVI) (Lobell et al. 2015), and normalized difference water index (NDWI) (Satir and Berberoglu 2016). Such indices provide regular (i.e., near-real-time, NRT) information on soil and crop vegetation status across large agricultural regions, and are used to assess crop status/condition, and to predict actual crop yield or as a yield proxy (Lobell et al. 2015). High quality data from weather and climate monitoring stations, field sensors and geospatial imagery obtained from Earth-orbiting satellites, provide a basis for improving crop monitoring and forecasting, and in turn, improving the robustness, effectiveness, and reliability of crop surveillance, irrigation, protection, and insurance programs. Moreover, payoffs of index products are more flexible, objective, and consistent than traditional approaches because they are based on the actual observation or realization of weather variables, e.g., rainfall and temperature driving changes in crop condition and yield. Traditional individual crop insurance policies are based on the reported crop production loss, which involves severe moral hazard, adverse selection as well as considerable costs for on-field verification of reported estimates and administration (Boyd et al. 2019). These benefits both for farmers and insurance companies have led to an increasing proliferation of index products in agricultural insurance, and the design of new weather indices and related insurance products plays a key role at the current nexus of digital agriculture and agricultural insurance. As result, precision agriculture and digital agriculture insurance are anticipated to substantially expand in the near future (see, for example, most recent 2021 the U.S. Department of Agriculture's National Institute of Food and Agriculture (USDA-NIFA) and the U.S. National Science Foundation (NSF) joint initiative for investing $220 million to 11 new NSF-led AI Research Institutes (USDA-NIFA 2021)). As such, we expect to see such AI tools to be more widely-adopted by farmers in monitoring and managing their crops.

In this paper we bridge the gap between DL and index-based agricultural insurance, that is, the area where DL tools have never been applied before. In particular, we discuss utility and limitations of DL, integrated with time-aware topological information on weather indices, to predict future crop yields. Our key modeling engine is the family of Long Short Term Memory (LSTM) networks, which is a special kind of recurrent neural networks (RNNs) for sequential data processing. LSTM solves vanishing long-term gradient problem by self-loops controlling flows of long duration. Since weather indices and, as a result, crop yields exhibit highly non-trivial nonseparable spatio-temporal dependence structure (by non-separability here we mean that spatial dependence among two locations depends on time, and vice versa), conventional (geo)statistical and ML approaches based on Euclidean distances might not adequately reflect the underlying hidden mechanisms behind formation and dynamics of crop status. To address this challenge, we introduce topological data analysis (TDA) and, in particular, time-aware topological signatures of weather indices constructed based on the tools of zigzag persistence to DL models for spatio-temporal evolution of crop yields. This allows us to simultaneously extract most salient shape properties of weather indices and crop yields which are invariant under continuous transformations such as twisting, bending, and compressing and which tend to consistently manifest over time and space resolutions. Despite its recent success in various domains neither TDA, nor zigzag persistence has ever been applied to agricultural studies or insurance analysis. We validate the proposed approach, namely, zigzag persistence-based LSTM (`ZZTop-LSTM`), in application to crop yield forecasting in Manitoba, Canada. Our experiments indicate that `ZZTop-LSTM` can deliver improved accuracy in forecasting crop vegetation status (i.e., crop condition) using satellite-based indices and shows promise in successfully accounting for higher-order interaction between crop yields, weather indices, and soil types (i.e., latent information which is not part of the model input). However, as most DL tools, `ZZTop-LSTM` tends to require longer observational samples compared to simpler models as decision trees. Overall, we find that upon securing a sufficiently long observational records, `ZZTop-LSTM` offers a promising direction for developing a more cost-effective, timely and accurate way to forecast crop status or condition across large agricultural regions, as a part of digital tools in agricultural insurance.

Related Work

Convolutional neural networks have been used to analyze more expensive locally sensed data by unmanned aerial vehicles (UAVs) (Nevavuori, Narra, and Lipping 2019) as well as less costly and globally accessible remote sensing data (Khaki, Pham, and Wang 2021) in agriculture. Furthermore, techniques such as Gaussian Process (You et al. 2017), transfer learning (Wang et al. 2018), and attention mechanisms (Lin et al. 2020) have been recently utilized to improve crop yield prediction accuracy and the associated model explainability. For instance, Jiang et al. (2019) report that 76% of heterogeneous regional variability of corn yield can be explained by RNN-LSTM model, outperforming both LASSO and RF approaches for least end-of-the-season yield estimation. Although there are many studies using ML in the agricultural insurance context, there yet exists none exploring integration of DL along with TDA for modeling complex relationships between crop yield and environmental variables.

Background

Topological Data Analysis and Zigzag Persistence TDA and, in particular, persistent homology (PH) assesses evolution of various shape patterns in the collected data as we vary a user-selected (dis)similarity threshold. By shape here we understand data properties which are invariant under continuous transformations. In case of point clouds in $\mathbb{R}$, as in our case, we start from constructing a distance graph based on the observed set and some suitable similarity measure.

Let $G = (V, E, \omega)$ be a representation of the dataset as the distance graph, where V is a node set, E is an edge set, and ω is an edge weight based on the user-selected (dis)similarity measure. Given an increasing sequence of (dis)similarity thresholds $\epsilon_1 < \epsilon_2 < \cdots < \epsilon_n$, we build a nested filtration of distance graphs such that $G_1 \subset G_2 \subset \cdots \subset G_n = G$. By equipping each distance graph with a combinatorial structure of abstract simplicial complex, we then count which topological patterns (e.g., connected components and 1-dimensional holes) appear and disappear as threshold ϵ changes, as well as record their lifespans. Topological features with longer lifespan are said to persist and are likely to contain valuable information on higher order interactions in the observed data. In turn, topological features with shorter lifespans are referred to as topological noise. As abstract simplicial complex, here we select a Vietoris-Rips (VR) complex due to its computational efficiency.

To extract topological signatures that manifest themselves over time, we can use the notion of zigzag persistence based on the quiver theory (Carlsson and de Silva 2010). Despite its promise for tracking time-dependent topological properties, applications of zigzag persistence not only in spatio-temporal processes but generally, in any domain beyond mathematics are yet nascent (Adams and Carlsson 2015; Chowdhury, Dai, and Mémoli 2018; Kim, Mémoli, and Smith 2020; Chen, Segovia-Dominguez, and Gel 2021). The key idea behind zigzag persistence is to set a threshold ϵ and then to consider a time-ordered sequence of simplicial complexes $\cdots \rightarrow VR(G_{i-1} \cup G_i, \epsilon) \leftarrow VR(G_i, \epsilon) \rightarrow VR(G_i \cup G_{i+1}, \epsilon) \leftarrow VR(G_{i+1}, \epsilon) \rightarrow \ldots$, where arrows correspond to addition or deletion of simplices indexed by time. That is, armed with the zigzag diagram, we can now track which topological features tend to persistently manifest over the sequence of time-ordered snapshots of the data, where each snapshot i is associated with $VR(G_i)$, $i \in \mathbb{Z}^+$ Topological signatures produced via zigzag persistence provide an alternative view of inherent time-conditioned shape characteristics of the process.

Lower panel of Figure 1 depicts a toy example of applying zigzag persistence on a dynamic network. Barcodes in the center and bottom diagrams trace dynamics of topological structures. Bottom barcode-diagram shows one long connected component, i.e. 0-dimensional feature, lasting for the whole time period, except at time $t = 3$ when the network splits into two connected components. Center barcode-diagram depicts the lifespan of holes, i.e. 1-dimensional features. The dynamic network mainly has two holes which appear and disappear at different timestamps. Notice that zigzag persistence homology tracks disappearance of lower hole at time $t = 3$ and posterior reappearance at time $t > 5$.

Long Short-Term Memory Networks LSTM consists of a chain like structure with repeating modules, where each unit is composed of a cell c_t and three transition functions, namely, input gate i_t, output gate o_t and forget gate f_t. The following equations present a forward pass $i_t = \sigma_i(W_i \cdot [h_{t-1}, x_t] + b_i)$, $o_t = \sigma_o(W_o \cdot [h_{t-1}, x_t] + b_o)$,$f_t = \sigma_f(W_f \cdot [h_{t-1}, x_t] + b_f)$, $g_t = \sigma_g(W_g \cdot [h_{t-1}, x_t] + b_g)$, $c_t = f_t \circ c_{t-1} + i_t \circ g_t$, $h_t = o_t \circ \tanh c_t$, where activation functions σ's introduce non-linearity to the linear forms of output vector from previous hidden layer h_{t-1} and current input data x_t. We update cell state c_t by forgetting part of previous memory and adding new information g_t. Cell information passes next LSTM unit through output gate.

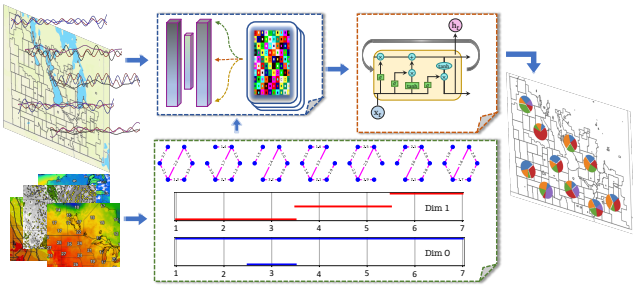

Figure 1: Architecture of `ZZTop-LSTM`.

Proposed Framework: `ZZTop-LSTM`

Our proposed Zigzag Persistence-based LSTM framework, `ZZTop-LSTM`, models the spatio-temporal dynamics of crop yields and introduces time-aware topological signatures of weather variables yielded by zigzag persistence, as complementary information into the LSTM model.

Forecasting Problem To predict yield and crop growth, we aim to discover hidden relationships in the data by inferring spatial and temporal patterns simultaneously. In this paper, we learn a function $\mathcal{F}(\cdot)$ that maps historical data $X_{t-p}, \ldots, X_{t-1}$ and geographical dynamic information $G_{t-p}, \ldots, G_{t-1}$ to future responses $Y_t, \ldots, Y_{t+h}$. We represent connection among locations, at time t, as a weighted undirected graph $G_t = (V, E, \omega_t)$, where V and E are the node set, respectively, and ω_t is an adjacency matrix with entries $\omega_t^{ij} > 0$. Here, p represents the windows size of historical signals and graphs, whilst h is the time ahead horizon.

Dynamic Network and Features Extraction Let $Z = \{Z_{ijt} : i = 1, \ldots, n; j = 1, \ldots, m; t = 1, \ldots, T\}$ be a multivariate time series of m weather variables over n locations of interest and time points t. We build a sequence of weighed undirected graphs $G_t = (V, E, w_t)$, where $w_t : E \rightarrow \mathbb{R}_{>0}$ defines time varying edge weights based on dissimilarity measures of weather variables Z_t. Connectivity of each graph is constructed in such a way that extracts essential spatio-temporal dependence structures at each week. For instance, municipalities in the same neighbourhood tend to be influenced by local weather events. Hence, we use geographic distance base on latitude & longitude via haversine formula (Winarno, Hadikurniawati, and Rosso 2017), and normalized Euclidean distance between weather variables to select edges in each network and assign its weights; where quantile-based parameter γ controls the number of edges. Hence, low values of γ help to reduce the computational burden of generating too many topological structures. Since we consider a sliding window approach, zigzag persistence is calculated on simplicial complexes of networks in the time window of size p via VR constructions, $VR(G_{t_k}, \epsilon) \rightarrow VR(G_{t_k}, G_{t_{k+1}}, \epsilon) \leftarrow VR(G_{t_{k+1}}, \epsilon) \rightarrow \cdots \leftarrow VR(G_{t_{k+p}}, \epsilon)$, which can be presented by barcode

plots and persistence diagram. Our feature extraction module computes summary statistics of birth, death and lifespan of zigzag barcodes at 0 and 1 dimensional features. As a result, we extend the concept of total persistence (Carrière et al. 2020) to the domain of zigzag persistence.

Capturing Temporal Dependencies LSTM is a standard tool to model temporal dependencies of multivariate time series with DL. Although LSTM is well suited to track a temporal structure of sequential features (Yu et al. 2019), the default LSTM architecture is not constructed to capture spatial information. Our proposed `ZZTop-LSTM` framework addresses this gap and introduces topological spatial information along with raw weather variables into LSTM. That is, the input series X_t consist of lagged weather variables Z_t as well as topological features U_t generated from corresponding window climatic dynamic networks $\{G_t\}$ with geographic connectivity constraints. Figure 1 shows a graphical representation of the main modules in our proposed `ZZTop-LSTM` framework. First, multivariate time series are used to construct a dynamic network. Second, we compute the zigzag persistence on the sequence of networks and extract summary statistics from zigzag barcodes. Third, agriculture monitoring variables, i.e. NDVI, weather variables and topological summaries serves as input to a series of three-stack LSTM layers. Finally, we obtain a weekly forecasting of NDVI values on each municipality of Manitoba. Therefore, `ZZTop-LSTM` complements conventional agriculture monitoring information with distinctive and persistent topological signatures, in order to introduce higher-order dependency properties which are otherwise inaccessible with conventional methods.

Computational Complexity To calculate the computational complexity of `ZZTop-LSTM`, we analyze the two main algorithmic components: Zigzag persistence and the LSTM architecture. Extraction of topological summaries from atmospheric variables depends upon applying a filtration function on simplicial complexes of time series of graphs. Let η be the number of simplices in the complex, and λ be the size of the largest simplicial complex. Current Zigzag persistence computation requires $\mathcal{O}(\lambda^2)$ memory space and the computational time complexity is $\mathcal{O}(\eta\lambda^2)$ (Carlsson, de Silva, and Morozov 2009). We alleviate the computational complexity through controlling the sparsity of graphs as described above. On the other hand, LSTM has time complexity per time step and weight $\mathcal{O}(1)$, hence, its complexity is $\mathcal{O}(\tau\omega)$, where ω and τ are the numbers of weights and time steps, respectively (Hochreiter and Schmidhuber 1997). Hence, the resulting total complexity of `ZZTop-LSTM` is calculated as the sum of Zigzag and LSTM complexities as $\mathcal{O}(\eta\lambda^2 + \tau\omega)$.

Experimental Results

Canadian Data Normalized difference vegetation index (NDVI) is widely used in agriculture literature, which reflects green plant coverage and health via plant photosynthetic activity. Green plants absorb photosynthetically active radiation (400-700 nm wavelengths) while re-emit near-infrared spectral region radiation (700–1300 nm wavelengths). Time-series data of normalized difference vegetation index (NDVI) derived from MODIS (Moderate Resolution Imaging Spectroradiometer) satellite imagery at the 250m resolution were acquired (Statistics Canada). The MODIS sensor's red and NIR channel ranges are 630–670 nm and 841–876 nm, respectively, and lie within the red and NIR spectrum bands of 600–700 nm and 700–1300 nm indicated earlier. The NDVI data is quality-controlled weekly data (i.e., 7-day composites, Monday to Sunday) during 2000–2018 (i.e., 19 growing seasons) from April to mid-October (or Julian weeks 14 (April 9–15) through week 41 (October 8–14)). The aggregation into weekly composites reduces error due to cloudy days among other error corrections required for the satellite imagery. A boundary file (LCSD00a16a ESRI shapefile) available from Statistics Canada is used to delineate the municipalities (counties) as census subdivisions. Centroids of each municipality (i.e., shapefile polygon) are references to obtain values of the climate variables for each region.

Daily climate data for the Province of Manitoba, Canada, were acquired from Daymet (https://daymet.ornl.gov/), providing spatially-interpolated estimates of daily maximum and minimum temperatures (°C), and precipitation (mm) at 1 km spatial grid resolution. Daily maximum and minimum temperatures are aggregated into weekly means, and daily precipitation is aggregated into weekly total precipitation, to correspond to the weekly timescale of the NDVI.

Experimental Settings `ZZTop-LSTM` is compared with the LSTM baseline model without topological signatures as well as statistical and ML methods which are currently the accepted benchmarks in agricultural practice, i.e., least absolute shrinkage and selection operator (LASSO), support vector regression with radial basis function kernel (SVR), and decision tree (DT) (You et al. 2017; Lin et al. 2020; Porth et al. 2020; Khaki, Pham, and Wang 2021). We use standard root mean square error (RMSE) and mean absolute error (MAE) as the main evaluation metrics in agriculture.

For the studied historical panel data (2000-2018), the last four years observations are reserved as test set. Since weeks in the crop growing seasons (14th week to 36th week) are of interest, the target is to predict 92 weeks ahead based on a month history information by setting number of lags $p = 4$. Zigzag persistence homology is computed from dynamic networks with time varying weights and connectivity over 37 municipalities of interest. We calculate topological signatures in the 434 sliding windows of size 4. Note that zigzag persistence homology is a pairwise operation and we can extract topological features from windows with arbitrary size over the same zigzag persistence of the whole time period. Hence, the input of training set contains 250 observations of four lagged sequences of underlying weather variables and topological summary statistics over the four-week window while output is corresponding a vector for 37 municipalities four years later.

Our zigzag persistence computations are based on Dionysus, which provides implementations of various persistent homology concepts[2]. Deep learning models are implemented in TensorFlow via Keras and those conventional approaches are developed with the scikit-learn library. Our source codes and datasets are online[3].

[2]Dionysus 2 https://mrzv.org/software/dionysus2/

[3]https://github.com/paper-code21

Model parameters The architecture of our LSTM model consists of three stacked LSTM layers with 256, 128 and 128 units, and a densely connected layer 37 outputs for considered municipalities or rural municipalities of Manitoba province, Canada. The input matrix consists of 4 weeks of weather variables and topological features aggregated, based on a sliding window. Zigzag persistence is determined from dynamic networks of the aggregated 4-week intervals of weather variables for each municipality with a connectivity threshold of 2/3. Topological features include summary statistics of barcodes on both ends of the sliding window in addition to those conventional topological signatures in general persistence homology analysis, which reflect some informative between-window dynamics of topological structure besides within-window characteristics. We used a 20% dropout probability to avoid overfitting, training the model using 8000 epochs and an early stopping condition of 800 epochs. Model performance is measured on the 92 weeks test set (growing seasons in 2015-2018).

Results Model validation statistics are summarized in Table 2 for a 4-year forecast window, 2015-2018 based on 10 runs due to randomness from initialization and dropout technique of deep model. The `ZZTop-LSTM` model reduces both the mean and variance in prediction error (i.e., lower RMSE and MAE), outperforming the LSTM model at significance level 10% and 5% (one-sided t-test), respectively. `ZZTop-LSTM` also improves over all other competitors currently accepted in agricultural practice.

The municipality-scale model comparison results are presented in Figure 2-Top. Overall, `ZZTop-LSTM` tends to consistently outperform all other benchmarks. `ZZTop-LSTM` is worse at the significance level of 10% than the baseline LSTM only in 3 counties (see Table 1). For conventional ML algorithms, SVR tends to be outperformed by `ZZTop-LSTM` at statistically significant level in all counties, and comparing to LASSO and decision trees, `ZZTop-LSTM` yields statistically significant improvement in the majority of counties, though spatial distribution of the results is more mixed. Gains in model accuracy for municipalities that contain very strongly to extremely calcareous (EC) parent matter (Figure 2-Bottom) are apparent. This needs to be further investigated. It is well known that calcareous soils have a high potential for crop yield increases (and high crop status) where adequate water and nutrients are available, so crop condition in such regions may be more resilient to weather and climate variability.

	`ZZTop-LSTM` versus			
Significance	LSTM	SVR	LASSO	DT
Improve at $< 1\%$ (***)	12	36	23	18
Improve at $1\% - 5\%$ (**)	5	1	0	1
Improve at $5\% - 10\%$ (**)	2	0	2	1
Improve at $> 10\%$	9	0	2	4
Decline at $> 10\%$	7	0	2	2
Decline at $5\% - 10\%$ (*)	3	0	0	0
Decline at $1\% - 5\%$ (**)	0	0	1	1
Decline at $< 1\%$ (***)	0	0	7	10

Table 1: Number of municipalities in which an improvement/decline is found using `ZZTop-LSTM` vs. its competitors.

Model	RMSE	MAE
`ZZTop-LSTM`	0.1482 ± 0.0017	0.1142 ± 0.0014
Competitors		
SVR-RBF	0.2034^{***}	0.1736^{***}
LASSO	0.1561^{***}	0.1243^{***}
Decision Tree	0.1519^{**}	0.1156^{**}
LSTM	$0.1508^{*} \pm 0.0019$	$0.1170^{**} \pm 0.0016$

Table 2: Validation statistics for benchmark and our `ZZTop-LSTM`, based on 4-year prediction window (2015-2018). Training data comprised 15 years starting from 2000.

Utility and Limitations

The topological LSTM approach is especially useful for deploying as a tool for forecasting crop status and yield. It offers a flexible method whereby assumptions regarding spatial dependencies, essential variables, different forecast windows and scenarios can be evaluated using big data streaming in from weather/climate stations, weather forecasts, and remote-sensing satellite data. Data at the field scale, in addition to the regional-scale, could also be potentially integrated from crop monitoring (e.g., drone imagery). While the current findings are promising, generalization of the new topological-based DL approach for operational crop yield forecasting and the associated insurance premium analysis require data from more regions and years for more comprehensive evaluation of all advantages and limitations, particularly, space-time uncertainty quantification of the derived forecasts. Also, additional explanatory variables (i.e. predictors of crop status and yield) need to be incorporated into the current model relating crop health and growth to environmental factors, such as soil water drainage, and important chemical and physical properties of different soil types. A well-drained soil retains water long enough for roots to absorb what the plant needs, and dries out sufficiently between rains or waterings so that roots can take up oxygen during high levels of soil moisture. Multiple interacting chemical, biological, and physical factors affect soil fertility, crop status and resultant crop yield. For example, soils can become acidic through rainfall and leaching, acidic parent material, organic matter decay, and harvest of high-yielding crops. Geospatial data for expanding the current study to include other important predictors is available through the Canadian Soil Information Service (CanSIS) as an authoritative source of soil data and land resource information for Canada.

Path to Deployment

The Government of Canada (Agriculture and Agri-Food Canada (AAFC) and Statistics Canada, StatCan) have developed and deployed a model-based, operational framework for forecasting the yield of major crops across Canada's agricultural area at a regional scale. This framework currently integrates weather, climate, and remote-sensing (i.e., EO) information (Newlands et al. 2014) and uses the Random Forest (i.e. decision tree) algorithm within a Bayesian modeling framework to forecast end-of-season crop yield. There is a critical need to improve the accuracy of such yield estimates and to extend the forecast window beyond a single growing season. Our findings reported here demonstrate that `ZZTop-LSTM` and, more generally, topologically-enhanced DL could improve yield estimation, compared to currently

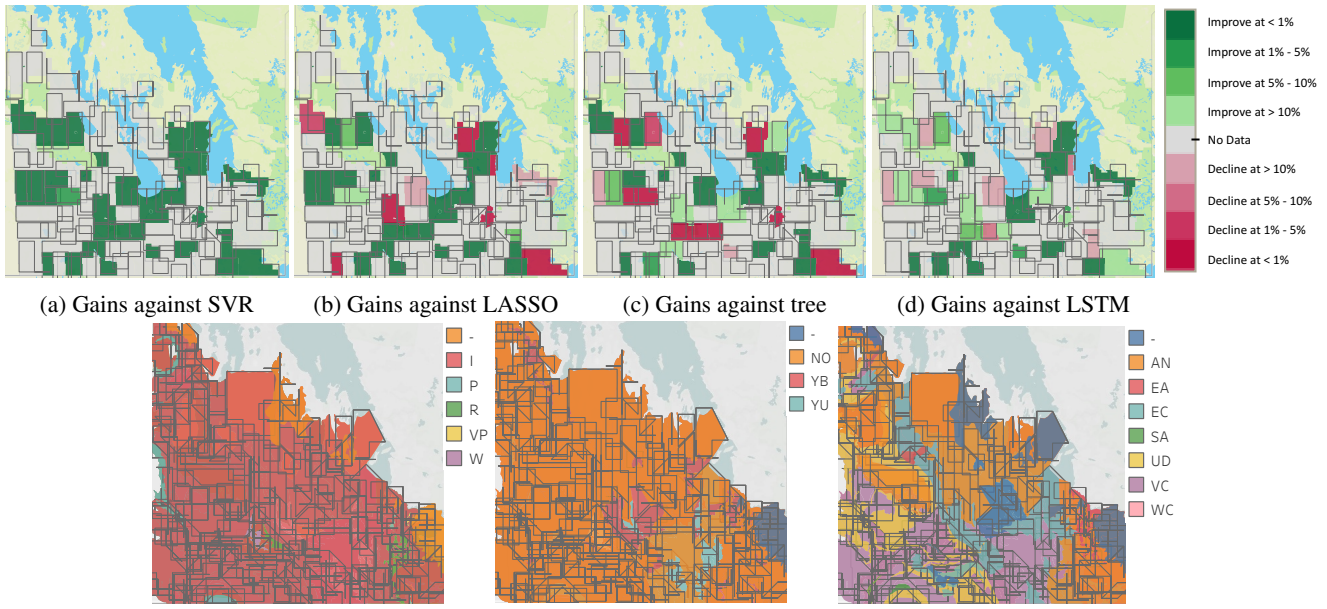

Figure 2: **[Top]** Regional-scale performance of Topological LSTM with respect to benchmarks. Improvements from Topological LSTM are shown in green, while declines are shown in red. Municipalities with no data are indicated as NA. The color bar on the right presents levels of significance. **[Bottom]** General physical and chemical characteristics for all of the soils identified in Manitoba's geographic region. **Left**. Soil drainage. Level: Very rapidly (VR), Rapidly (R), Well (W), Moderately well (MW), Imperfectly (I), Poorly (P), Very poorly (VP), Not applicable (-). **Center**. Water. Level: Always (YB), Growing season (YG), Non growing season (YN), Unspecified period (YU), Never (NO), Not applicable (-). **Right**. Parent material chemical property. Level: Undifferentiated (UD), Extremely / Strongly Acidic (EA), Medium Acid to Neutral (AN), Weakly Calcareous (WC), Moderately / Very Strongly Calcareous (VC), Extremely Calcareous (EC), Calcareous and Saline (SA), Not Applicable (-)

deployed methodology, while also providing 4-year look-ahead forecasts (with associated uncertainty) for guiding agricultural policy and insurance rate-making decision making under climate change. By incorporating spatial and temporal dependency between crops and their environment, the exposure and vulnerability of different crops to extreme weather and a changing climate conditions can be better represented and accounted for. Future work will seek to further validate and evaluate the Topological LSTM method within the existing Canadian Crop Yield Forecaster (CCYF) operational forecasting framework. The path to deployment will require several key steps: 1) validation using multi-spectral satellite indices and optimization for different crop types, 2) integrating Canada's crop inventory geospatial data providing high-resolution delineation of crop types over time, 3) uncertainty quantification and its impact on forecasting under different simulated extreme weather events and future climate scenarios, 4) integration of model-based weather forecast scenario output available out to 14 days (open data provided by Canadian Meteorological Service (MSC) of Environment and Climate Change Canada, ECCC), and 5) actuarial rate-making analysis to evaluate pricing mechanisms and risk premium payouts. In the future, it could become the method deployed in this national operational tool used by agricultural stakeholders (e.g., farmers, crop advisors, policy analysts, crop insurers/re-insurers). More generally, for better implementation, it is important to coordinate efforts among all stakeholders, i.e., insurance and reinsurance companies, agriculture departments and regulatory bodies, as well as institutions collecting and distributing weather and climate data. This pilot project is the first step of the broader AAFC research initiative to unify such efforts.

Conclusions

Our findings for wheat in Canada (Manitoba), support those highlighted by Jiang et al. (2019) on the potential of an LSTM approach as a methodology for operational prediction of regional-scale crop yield. Scaling up the current study by applying `ZZTop-LSTM` in Canada using a larger data set would enable expanded training and validation of this approach. This is similar to the data requirements of other DL methods for scaling up. The `ZZTop-LSTM` model generates greater accuracy estimating crop status but with less variance. This has implications for crop insurance, as it offers reduced variance in estimating crop status and end-of-season yield. This, in turn, reduces basis risk in crop insurance, thereby offering a more reliable estimate of a farmer's actual crop losses in weather- and/or NDVI-based index insurance schemes. Indeed, more accurate pricing resulting from DL tools similar to `ZZTop-LSTM` facilitates low sum insured and low rate products with online purchase and automatic settlement, especially for small size farms, hence, promoting regional economic development, especially, in developing countries.

Acknowledgments

Newlands was supported by the Canadian Agricultural Partnership (CAP) Program of Agriculture and Agri-Food Canada (AAFC) Project No. 1587 (J-001387.001.11). This work is also supported by Casualty Actuarial Society (CAS) CKER, NSF DMS 1925346, NSF ECCS 1824716 and NASA 80NSSC20K1579 grants.

References

Adams, H.; and Carlsson, G. 2015. Evasion paths in mobile sensor networks. *Int. J. Robot. Res.*, 34(1): 90–104.

Boyd, M.; Porth, B.; Porth, L.; and Turenne, D. 2019. The Impact of Spatial Interpolation Techniques on Spatial Basis Risk for Weather Insurance: An Application to Forage Crops. *NAAJ*, 23(3): 412–433.

Carlsson, G.; and de Silva, V. 2010. Zigzag Persistence. *FoCM*, 10(4): 367–405.

Carlsson, G.; de Silva, V.; and Morozov, D. 2009. Zigzag Persistent Homology and Real-Valued Functions. In *SCG*, SCG '09, 247–256. New York, NY, USA: ACM.

Carrière, M.; Chazal, F.; Glisse, M.; Ike, Y.; and Kannan, H. 2020. Optimizing persistent homology based functions. *arXiv: 2010.08356v2*.

Chen, Y.; Segovia-Dominguez, I.; and Gel, Y. R. 2021. Z-GCNETs: Time Zigzags at Graph Convolutional Networks for Time Series Forecasting. In *ICML*.

Chowdhury, S.; Dai, B.; and Mémoli, F. 2018. The importance of forgetting: Limiting memory improves recovery of topological characteristics from neural data. *PloS one*, 13(9): e0202561.

Guo, H.; Woodruff, A.; and Yadav, A. 2020. Improving lives of indebted farmers using deep learning: Predicting agricultural produce prices using convolutional neural networks. In *AAAI*, volume 34, 13294–13299.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural Comput.*, 9(8): 1735–1780.

Jia, X.; Wang, M.; Khandelwal, A.; Karpatne, A.; and Kumar, V. 2019. Recurrent Generative Networks for Multi-Resolution Satellite Data: An Application in Cropland Monitoring. In *IJCAI*.

Jiang, H.; Hu, H.; Zhong, R.; Xu, J.; Xu, J.; Huang, J.; Wang, S.; Ying, Y.; and Lin, T. 2019. A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the US Corn Belt at the county level. *Glob. Change Biol.*, 26(3): 1754–1766.

Khaki, S.; Pham, H.; and Wang, L. 2021. Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. *Sci. Rep.*, 11(1).

Kim, W.; Mémoli, F.; and Smith, Z. 2020. Analysis of dynamic graphs and dynamic metric spaces via zigzag persistence. In *Topological Data Analysis*, 371–389. Springer.

Lin, T.; Zhong, R.; Wang, Y.; Xu, J.; Jiang, H.; Xu, J.; Ying, Y.; Rodriguez, L.; Ting, K. C.; and Li, H. 2020. DeepCropNet: a deep spatial-temporal learning framework for county-level corn yield estimation. *Env. Res. Letters*, 15(3): 034016.

Lobell, D. B.; Thau, D.; Seifert, C.; Engle, E.; and Little, B. 2015. A scalable satellite-based crop yield mapper. *Remote Sens. Environ.*, 164: 324–333.

Nevavuori, P.; Narra, N.; and Lipping, T. 2019. Crop yield prediction with deep convolutional neural networks. *Comput. Electron. Agric.*, 163: 104859.

Newlands, N. K.; Zamar, D. S.; Kouadio, L. A.; Zhang, Y.; Chipanshi, A.; Potgieter, A.; Toure, S.; and Hill, H. S. J. 2014. An integrated, probabilistic model for improved seasonal forecasting of agricultural crop yield under environmental uncertainty. *Front. Environ. Sci.*, 2: 17.

Porth, C. B.; Porth, L.; Zhu, W.; Boyd, M.; Tan, K. S.; and Liu, K. 2020. Remote Sensing Applications for Insurance: A Predictive Model for Pasture Yield in the Presence of Systemic Weather. *NAAJ*, 24(2): 333–354.

Quarmby, N. A.; Milnes, M.; Hindle, T. L.; and Silleos, N. 1993. The use of multi-temporal NDVI measurements from AVHRR data for crop yield estimation and prediction. *Int. J. Remote Sens.*, 14(2): 199–210.

Satir, O.; and Berberoglu, S. 2016. Crop yield prediction under soil salinity using satellite derived vegetation indices. *Field Crops Res.*, 192: 134–143.

Shoshi, H.; Hanson, E.; Nganje, W.; and SenGupta, I. 2021. Stochastic Analysis and Neural Network-Based Yield Prediction with Precision Agriculture. *J. of Risk and Fin. Management*, 14(9): 397.

USDA-NIFA. 2021. USDA-NIFA and NSF Invest $220M in Artificial Intelligence Research Institutes. *USDA Press Release*.

van Klompenburg, T.; Kassahun, A.; and Catal, C. 2020. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.*, 177: 105709.

Wang, A. X.; Tran, C.; Desai, N.; Lobell, D.; and Ermon, S. 2018. Deep Transfer Learning for Crop Yield Prediction with Remote Sensing Data. In *Proc. of the 1st ACM SIGCAS COMPASS*. ACM.

Winarno, E.; Hadikurniawati, W.; and Rosso, R. N. 2017. Location based service for presence system using haversine method. In *ICITech*, 1–4.

Worrall, G.; Rangarajan, A.; and Judge, J. 2021. Domain-guided Machine Learning for Remotely Sensed In-Season Crop Growth Estimation. *arXiv: 2106.13323v1*.

You, J.; Li, X.; Low, M.; Lobell, D.; and Ermon, S. 2017. Deep gaussian process for crop yield prediction based on remote sensing data. In *AAAI*, 4559–4565.

Yu, Y.; Si, X.; Hu, C.; and Zhang, J. 2019. A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures. *Neural Comput.*, 31(7): 1235–1270.

Zhu, W.; Porth, L.; and Tan, K. S. 2019. A credibility-based yield forecasting model for crop reinsurance pricing and weather risk management. *Agr. Fin. Rev.*, 79(1): 2–26.