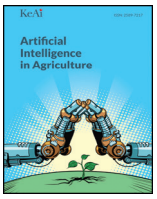
Artificial Intelligence in Agriculture 5 (2021) 208–222

Contents lists available at ScienceDirect

Artificial Intelligence in Agriculture

journal homepage: http://www.keaipublishing.com/en/journals/artificial-intelligence-in-agriculture/



Combining machine learning, space-time cloud restoration and phenology for farm-level wheat yield prediction

Andualem Aklilu Tesfaye [a,*], Daniel Osgood [b], Berhane Gessesse Aweke [c]

[a] Department of Remote sensing at Ethiopian Space Science and Technology Institute, Addis Ababa University, Ethiopia
[b] International Research Institute for Climate and Society, Earth Institute, Columbia University, United States of America
[c] Department head of the remote sensing unit at Ethiopian Space Science and Technology Institute, Addis Ababa University, Ethiopia

ARTICLE INFO

Article history:
Received 17 August 2021
Received in revised form 8 October 2021
Accepted 9 October 2021
Available online 12 October 2021

Keywords:
Cloud restoration
Ensemble learning
Machine learning
Phenology
Sentinel-2
Wheat yield prediction

ABSTRACT

Though studies showed the potential of high-resolution optical sensors for crop yield prediction, several factors have limited their wider application. The main factors are obstruction of cloud, identification of phenology, demand for high computing infrastructure and the complexity of statistical methods. In this research, we created a novel approach by combining four methods. First, we implemented the cloud restoration algorithm called gapfill to restore missed Normalized Difference Vegetation Index (NDVI) values derived from Sentinel-2 sensor (S2) due to cloud obstruction. Second, we created square tiles as a solution for high computing infrastructure demand due to the use of high-resolution sensor. Third, we implemented gapfill following critical crop phenology stage. Fourth, observations from restored images combined with original (from cloud-free images) values and applied for winter wheat prediction. We applied seven base machine learning as well as two groups of super learning ensembles. The study successfully applied gapfill on high-resolution image to get good quality estimates for cloudy pixels. Consequently, yield prediction accuracy increased due to the incorporation of restored values in the regression process. Base models such as Generalized Linear Regression (GLM) and Random Forest (RF) showed improved capacity compared to other base and ensemble models. The two models revealed RMSE of 0.001 t/ha and 0.136 t/ha on the holdout group. The two models also revealed consistent and better performance using scatter plot analysis across three datasets. The approach developed is useful to predict wheat yield at field scale, which is a rarely available but vital in many developmental projects, using optical sensors.

1. Introduction

Crop yield estimation is at the hub of numerous global developmental challenges such as food security, hunger reduction, and international crop trade among others (Kim et al., 2019; Dodds and Bartram, 2016). Yield estimation is the process of accurately predicting crop yield before harvest. Remote sensing technologies coupled with empirical methods used to predict crop yield having some comparative advantages, notably, covering a large area, enabling continuous monitoring, and providing timely data (Potgieter et al., 2014; Lobell et al., 2015). These methods discriminate the potential of a range of vegetation indices (VIs). The relationship between satellite-derived VIs and crop yield is established because important crop characteristics that determine crop yield such as leaf area index, biomass, and chlorophyll content are represented by VIs. Specifically, VIs record canopy level properties (reflectance, transmission, and absorbance) which enable to infer valuable canopy attributes (biochemical, physiological, and morphological) (Zhao et al., 2020).

Nonetheless, several issues have challenged the application of VIs for crop yield prediction. Three major categories could be identified: sensor (obstruction of images by cloud, missing orbits, and geometry artifacts), data (absence of long years of crop yield data), and statistics (the complex relationship between predictor and response variables). Besides, images obtained at some crop growth stages, which are commonly referred to as critical stages, have paramount importance for crop yield prediction. A study revealed that using Quickbird image at end of the heading and in particular after inflorescence fully emerged, NDVI is highly correlated with yield with an average correlation coefficient of 0.77 (Kumhálová and Matějková, 2017). However, in some studies due to cloud coverage, observations of greenness measured by NDVI at important critical windows such as flowering and grain filling stages were left out from the regression process (Kumhálová and Matějková, 2017; Zhao et al., 2007).

Although several solutions are available, cloud restoration (gap filling) methods are among the most widely applied ones. Cloud

* Corresponding author.
E-mail addresses: anduak@yahoo.com (A.A. Tesfaye), deo@iri.columbia.edu (D. Osgood).

https://doi.org/10.1016/j.aiia.2021.10.002

restoration methods employ some sort of algorithms for replacing cloudy pixels with estimates from the non-cloud part of the image. In terms of exploiting sensor potential, three major categories of cloud restoration exist: temporal, spatial, and spatio-temporal (Gerber et al., 2018). Methods under the temporal category include most of the existing algorithms. For instance, the widely established method called TIMESAT exploits the temporal dependence structure and smooth time series data and overlooks spatial dependence (Verger et al., 2013; Atkinson et al., 2012; Hird and McDermid, 2009). In general, spatio-temporal methods are rarely available and in this regard, one of the recently available methods named gapfill exploits the spatial coherence and temporal regularity of images (Garonna et al., 2016; Zscheischler and Mahecha, 2014). In terms of the root-mean-squared error (RMSE), it overperformed two of its competitors: the gapfill-MAP and TIMESAT (Weiss et al., 2014). As it is available in R platform, it is open access and transparent, and it could easily be applied and further developed (Gerber et al., 2018). Nonetheless, gapfill was applied using MODIS (Moderate Resolution Imaging Spectroradiometer) NDVI which has medium spatial resolution and its performance on high spatial resolution is left unaddressed.

The availability of high spatial and temporal remote sensing products offered untapped potential in many applications. Nonetheless, as these products are both data and compute intensive, their usage has its challenges (Zhong et al., 2013).

Voluminous pixel and multiple spectral bands (multispectral/hyperspectral) of large size satellite data cause great challenges to read, process, store and display (Pektürk and Ünal, 2018). The processing challenge of high volume remote sensing data is associated with the intrinsic nature of data mining algorithm. Specifically, the computational complexity of processing algorithms is superliner to the size of the sample and the number of samples (Kalluri et al., 2001). This is more significant in implementing spatio-temporal algorithms that discriminate both the high temporal and spatial resolution of newly available sensors. One solution has been the use of High Performance Computing (HPC) that includes clusters, grids, clouds, distributed networks or specialized hardware devices that provide useful architectural developments for increased computation of remote sensing images (Lee et al., 2011). For instance, a spatio-temporal image restoration algorithm, gapfill, demanded 80 cores running for 10 h using MODIS sensor, which is only high temporal (Gerber et al., 2018). Likewise, a HiTempo platform, which is designed to facilitate the analysis of time-series satellite images under parallelized setup, required 32 processors to process 415 tiles over 9 years (Van Den Bergh et al., 2012). However, due to their recent advancement and hence limited availability, HPCs are rarely available in many image processing environments.

To model, the complex relationship between crop yield and potential predictors various statistical and machine learning (ML) approaches used. Correlation analysis was less useful for understanding and quantifying yield response (Drummond et al., 2003). Several multiple linear regression models were widely applied and overall revealed poor results (Drummond et al., 1995; Kbakural et al., 1999). Nonetheless, other authors reported linear models performed well on some datasets (Kbakural et al., 1999). In some studies, polynomial methods showed some improvement over strictly linear models (Kitchen et al., 1999). Nonlinear methods such as boundary line analysis (Kitchen et al., 1999), state-space analysis (Wendroth et al., 1999), bayesian networks, and regression trees (Adams et al., 1999) were also used. Nonetheless, the implementation and the lack of common measurement scale for outputs from these diverse models were the major difficulties. Artificial Neural Network (ANN) model showed higher accuracy than multivariate regression models using Advanced Very High Resolution Radiometer (AVHRR) sensor for winter wheat (Jiang et al., 2004). On other studies, Support Vector Machine used to estimate rice yields (Jaikla et al., 2008). Therefore, the selection of a given method needs considering the characteristics of the predictors and response variable. Nonetheless, based on the current trend, machine learning methods are not only contemporary but also preferred ones.

Satellite-based farm-level estimation methods are seldom available. This is associated with the widely available and most exploited sensors (MODIS and Landsat) possess only limited potential to be applied at finer spatial scales. In this regard, the launch of Sentinel 2 (S2) with high spatial and temporal resolution has offered a new opportunity. A recent study compared the commercial Skysat, RapidEye satellites, and the publicly accessible S2 using VIs for predicting farm-level crop yield in the Eastern African region. Though the study overlooks the crop phenology, it evaluates the link between farm size units with model performance (Gomez et al., 2019). The study applied some machine learning models on yield predictor variables derived from S2 to predict potato yield. The study reported that potato yield is predicted 1–2 months ahead of the harvest period with satisfactory result (11.16% RMSE, $R^2 = 0.89$ and 8.71% MAE) (Gomez et al., 2019). Similarly, in another study, the S2 sensor integrated with the ecosystem model used for estimating cotton yield over the Southern United States (He and Mostovoy, 2019).

Therefore, this study motivated to address some of the key challenges of farm-level crop yield estimation using optical sensors. This study is novel as it integrates and implements four methods: application of gapfill on high-resolution S2 sensor, application of gapfill following crop phenology stage, the use of tilling approach within gapfill framework, and evaluation of observations from cloud restored images for yield prediction using seven machine learning and ensemble methods. We hypothesized that by discriminating the spatio-temporal potential of the S2 sensor, the gapfill method will provide a good estimate of cloudy pixels and the machine learning methods will learn the non-linearity between the NDVI variable and winter wheat yield. Hence, the two methods: gapfill and machine learning potentially complement and offer an increased skill for wheat yield prediction. Therefore, the purpose of this study is twofold. First, we implement the gapfill algorithm for cloud restoration on NDVI variables derived from high spatial and temporal resolution S2 sensor. Second, observations from restored images combined with original (from cloud-free images) values used for wheat yield prediction.

2. Methodology

2.1. Study area, preprocessing and dataset description

Wheat is one of the widely produced crops in Ethiopia. It accounts for 16% of the total grain production in which 4.8 million farmers take part. The current study is placed in some parts of the wheat belt region of Ethiopia and in particular in Arsi-Sire district which is located in Central-Eastern Oromia Region. Rainfed smallholder crop production dominated in the study area, where wheat monocropping exercised. Annually the area has two crop growing seasons, the major season called: 'Meher' (September–February) and the second 'Belg' (March–August) (Brasesco et al., 2019).

We used the multispectral bands of S2 sensors with 10 m spatial and 5 days temporal resolutions. The original accessed S2 product was level-1C data (top of atmosphere) downloaded from the Copernicus Open Acess Hub. We used Sentinel Application Platform (SNAP) and R-programming software to implement the preprocessing activities and to prepare the final input data for the gapfill process. First, an atmospheric correction algorithm implemented to get level-2A product (bottom of the atmosphere) from level-1C. Second, the NDVI product derived from a level-2A product. Third, level-2A products resampled to get quality scene classification output. Fourth, the final NDVI mask product was prepared by masking some classes (dark feature shadow, cloud shadow, cloud medium probability, cloud high probability, and thin cirrus) of the quality scene classification product. Finally, the mask applied to NDVI data (Fig. 1).

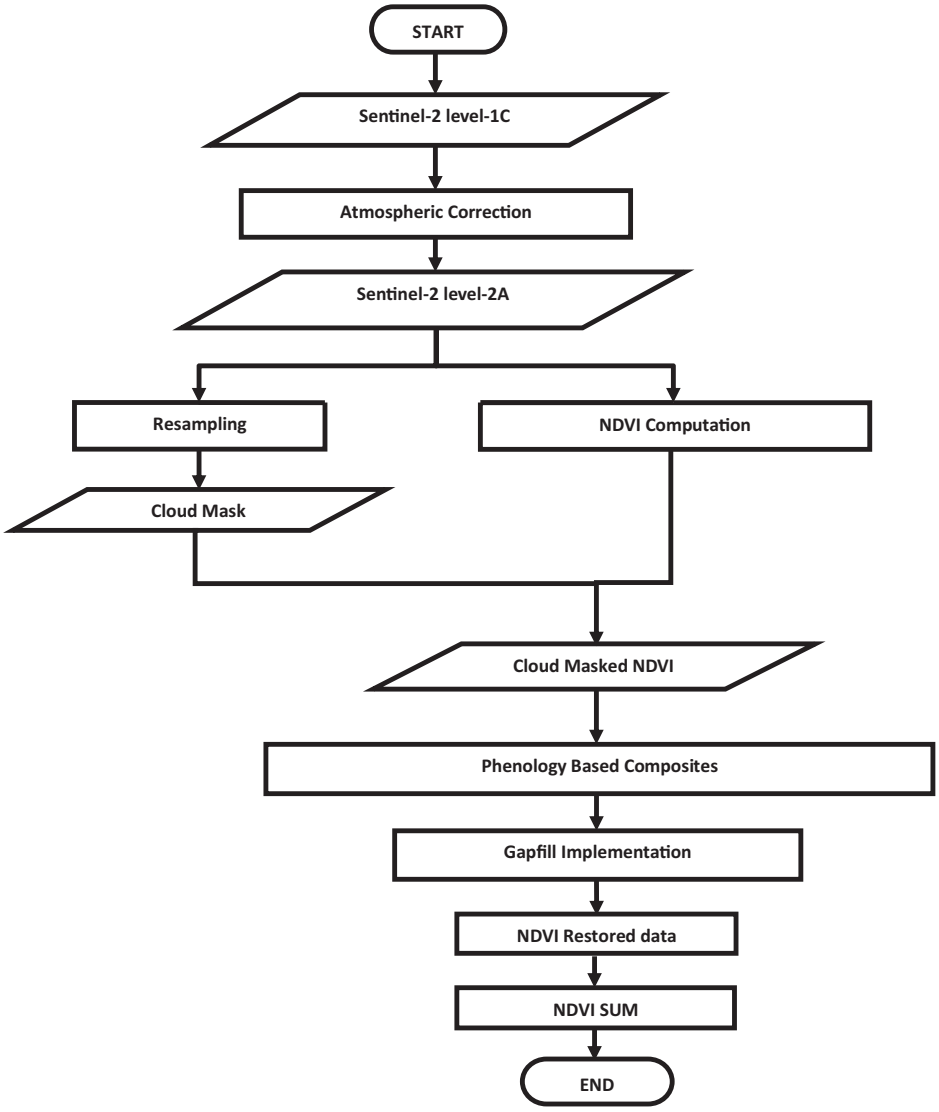


Fig. 1. Data processing workflow from image preprocessing to cloud restoration.

Based on previous studies, images obtained at some crop stages demonstrated increased potential for yield estimation. For cereal crops, these stages range from heading up to ripening (Fischer, 2011). Likewise, Mirasi et al., 2019 reported milky and ripening stages as key periods. Under the current study area, planting enters as early as July 8 and stays up to July 18; and harvesting starts on October 29 and sustains up to November 9. Therefore, this study focused on the major crop-growing period: 'Meher', and specifically on the crop development stage from the date of September 26 – October 26 which referred as study season. From a phenology perspective, the study season overlaps with crop development stages such as Flowering, Anthesis, Development of fruit, and Ripening.

The input dataset for gapfill should be prepared in four-dimensional arrays: x (dimension one), y (dimension two), s (season: dimension three), and a (year index: dimension four). The input dataset in this study has a spatial extent of 1486 grids along the x and 1837 grids along the y. To the season index, seven images found in the study season used. We grouped images per phenology considering the hypothesis that images found in one phenology stage represents a single physiological crop development and could yield a composite that offers reliable prediction values. The study applied data from three years: 2017, 2018, and 2019, which are year indexes. Finally, the total number of images used in the restoration process equals 21 (seven images per season for each year considered).

The key field-collected data in this study was dry weighted wheat yield measured in ton per hectare (t/ha). It was collected from 67 farmers' fields for two crop seasons: 2017 and 2018. Furthermore, in this study, with the primary purpose of getting a better yield prediction method, some potential predictors added. Specifically, two types of fertilizers that have been widely used in the study area such as Urea fertilizer that contains 46% nitrogen; and NPS, which contain nitrogen, phosphorous and sulfur with the ratio of 19% N, 38% P2O5 and 7% S were used (Meessen and Petersen, 2010).

2.2. Gapfill description

The overall workflow of the gapfill method could be presented in two major parts: the extraction (subset) component and prediction component, which were implemented in sequential order. In R platform, we used packages that include gapfill, raster, foreach, doParallel, and rgdal to support the overall process. Details of the gapfill method are described in the paper (Gerber et al., 2018).

2.2.1. Extraction component

This is the first step in which the algorithm iteratively selects a big enough neighborhood which is referred to as a prediction set along the spatio-temporal dimension. The subset is constructed around the target value (missing value) considered at a time. The row remote sensing data need to be stored in a four-dimensional array and missing values will be stored as NA. For an array Z with four dimensions x (x spatial dimension), y (y spatial dimension), s (season or position within a year), and a (year index) for a target value T, the subset function selects a suitable subset of Z around T. The subset function is given by the following equation:

$$\begin{aligned} f\,(Z,i) &= f\,(Z;i,xT,yT,sT,aT,\lambda x,\lambda y,\lambda s,\lambda a) \\ &= Z[(xT-(\lambda x+i)):(xT+\lambda x+i),(yT-(\lambda y+i)):(yT+\lambda y+i), \\ &\quad (sT-\lambda s):(sT+\lambda s),(aT-\lambda a):(aT+\lambda a)]. \end{aligned} \tag{1}$$

$\lambda x, \lambda y, \lambda s, \lambda a \in N$ are tuning parameters, which define the initial size of the subset; and i that grows the subset in the spatial dimension, will have values from 0 and iteratively increases by 1 when the subset does not contain enough data.

2.2.2. Prediction component

2.2.2.1. Decision on the subset. This is the first step of the prediction component that decides whether it is possible to predict the target value defined by Eq. (1) above. The process evaluates two criteria; first, the subset should have the minimum number of non-empty sub-images defined by $\theta 1$. Second, the sub-images in the subset containing the target value should have the minimum number of observed values in the target sub-image set by $\theta 2$. If the two criteria met, then the specific subset is selected as prediction set. A sub-image is any image found within the array dataset identified by season and year indexes.

2.2.2.2. Rank the image. At this stage, for all the sub-images within the prediction set, a scoring algorithm ranks each of them. The score of a sub-image is the proportion of values in the sub-image that are larger compared to the values at the same spatial coordinates in all other sub-images. The sub-images are ranked by increasing score, i.e., a sub-image with the smallest score will be assigned rank 1 and so on.

2.2.2.3. Estimate quantile. For sub-images within the prediction set having an observed value at the spatial location of the target pixel, here, the algorithm determines to which empirical quantile level that value corresponds relatively to all values of the image. The observed values in the prediction set are used to estimate an empirical cumulative distribution function for each sub-image. The final quantile level will be the mean of all quantiles levels for all the sub-images in the prediction set. A tuning parameter, v, which is the minimum number of quantiles, controls the process.

2.2.2.4. Quantile regression. The final step of the prediction component is development of a quantile regression model and applying it for prediction. An intercept and the associated rank, r, is added as inputs in the regression process as linear predictors for all observed values in the prediction set. For a given quantile level, τ, the regression equation is given by.

$$Q(\tau \mid r) = \beta 0(\tau) + \beta 1(\tau) r \tag{2}$$

Where $\beta 0(\tau)$ and $\beta 1(\tau) \in R$ are the coefficients. Then, the prediction of the target value is implemented based on fitted quantile regression.

2.2.3. Validation of the prediction procedure

Validation is the process of evaluation of outputs from the prediction process compared to observed values. The gapfill algorithm enables the computation of RMSE using a cross-validation procedure. The procedure starts with true data and removes some validation points to obtain artificially generated points, which referred to as data observed. Then, the observed data will be predicted and filled with values. The comparison between data filled and the original data set will yield the metric RMSE (Gerber, 2018). In addition, the validity of the restored dataset assessed visually employing spatial coherence and a non-artificial (natural) look.

2.3. Machine learning methods

In this study three groups of data mining techniques: linear regression, individual machine learning (base models), and ensemble methods applied. Linear regression methods, both linear and multiple, are used to evaluate the influence of incorporating restored values from gapfill on wheat yield prediction. The latter two groups of models applied to establish prediction models between predictors and response variable. In the second group, seven types of machine learning methods i.e., Regularized Regression (GLMNET), Generalized Linear Regression (GLM), Neural Network (NNET), K-nearest Neighbors (KNN), Recursive Partitioning and Regression Trees (RPART), Support Vector Machine (SVM) and Random Forest (RF) applied. In general, the potential of machine learning methods roots in the use of large dataset for training; nonetheless, as this study used a small dataset (67 points) we applied seven different machine learning methods in light of harnessing the potential of each method under small dataset domain. Besides, the seven methods cover the wide diversity and potential representing various groups of ML in terms of their function. Thus, KNN and SVM methods are selected from instance-based algorithms (Aha et al., 1991) whereas GLMNET represents the regularization algorithms group. RPART is selected from decision tree algorithms category while RF is obtained from ensemble group of algorithms. Neural Net has its own category and GLM model is from the group of linear regression (Piccini, 2019; Brownlee, 2016).

For each of the seven models, the following Sections (2.3.1–2.3.7) presented a short description of the methods, the hyperparameters with their tuned values as well as the final selected hyperparameters.

2.3.1. Regularized regression (GLMNET)

It fits generalized linear and similar models via penalized maximum likelihood using gaussian regression model (Friedman et al., 2010). It computes lasso or elastic net penalties at a grid of values for the regularization parameter lambda. Regularization methods are applied to estimate dependable predictor coefficients when predictors are correlated. All the predictors in the model are kept if ridge regression is used. LASSO ensures sparsity of the results by shrinking some coefficients exactly to zero, while Elastic Net is a hybrid of ridge regression and LASSO by adjusting the values of hyperparameter α (alpha) (Friedman et al., 2009) that has values in the range of $0 \le \alpha \le 1$. An alpha = 1 is the lasso penalty, and an alpha = 0 applies the ridge penalty. In this study, α is searched at three values of 0, 0.2, and 0.05, while lambda, which controls the amount of regularization, is tuned for values of 10^seq (−3,−2, length = 8). The study finds an alpha = 0, and lambda = 0.007196857 as best tuning parameters.

2.3.2. Support vector machine (SVM)

Support Vector Machine is a supervised non-parametric algorithm characterized by using kernels, which act, on the margins (Gunn, 1998). In regression, the input is mapped to a high-dimensional feature space using a kernel function. Then, a linear regression model is constructed in the new feature space (Cai et al., 2019; Hearst et al., 1998). In this study, the radial basis kernel, which is one of the widely used kernels, that fits non-linear model used. Then, parameters such as sigma and cost (C) are searched at values of 2^ (−11:-9) and 2^ (5,7) respectively. The study finds, sigma = 0.0004882812 and C = 128 as best tuning parameters.

2.3.3. K-nearest neighbors (KNN)

K-nearest Neighbors (KNN) is a supervised machine learning algorithms used for classification and regression. In KNN, the number of neighbors (k) is a constant used to calculate nearest neighbors distances vector (RSrivastava, 2020). KNN is an instance-based learning; the output is the mean of the values of its k nearest neighbors (Appelhans et al., 2015). The study tuned k for the range of 1:16 and selected k = 12 as the final value.

2.3.4. Neural network (NNET)

It is a network model consisting of input, hidden, and output layers to emulate a biological neural system. This study fits a single-hidden-layer neural network. In NNET, the size parameter assigns the number of units in the hidden layer. The weight decay parameter, or regularization, is a technique applied to the weights of a neural network. This parameter prevents the weight from increasing too large as it multiplies the weights after each update. In this study, linear activation function is applied. The size parameter is searched for size = 1:3 and decay = 10^ (seq(−3,−1, length = 5)) (Bishop, 1995; Fritsch et al., 2019). The NNET implementation in this study obtained size = 1, decay = 0.1, trace = FALSE, and linout = TRUE as final best values.

2.3.5. Random forest (RF)

It aggregates the predictions made by multiple decision trees. Each tree relies on the values of a random vector sampled independently and with the same distribution (Breiman, 2001). The most important parameter, mtry that is the number of predictors that are picked randomly is searched for mtry = 2:6 (Breiman, 2002). The Random Forest implementation in this study obtained mtry = 2 as best value.

2.3.6. Recursive partitioning and regression trees (RPART)

These are decision tree algorithms used for classification and regression learning tasks (Breiman et al., 2017; Terry and Beth, 2019). The complexity parameter (cp), which is the minimum improvement in the model needed at each node, imposes a penalty to the tree for having too many splits. It helps to select the optimal tree size as it controls the size of the decision tree (Therneau and Atkinson, 2019). The study tuned cp at values of 10^seq (−3,−0.75, length = 25). The study finds an optimum cp = 0.07498942.

2.3.7. Generalized linear regression (GLM)

It generalizes linear regression by allowing the linear model to be related to the response variable via a link function by allowing the magnitude of the variance of each measurement to be a function of its predicted value (Nelder and Wedderburn, 1972; Hardin et al., 2007). This model allows us to build a linear relationship between the response and predictors, even though their underlying relationship is not linear. The error distribution of the response variable need not to have normal distribution, however it assumed to follow an exponential family of distribution (i.e. normal, binomial, poisson, or gamma distributions). It applies a unique link function for each of the probability distribution (McCullagh and Nelder, 1989). Thus, this study used a gaussian probability distribution with identity link function.

The dataset that contains both predictors and response variables was partitioned into 80:20 ratios for training and testing (holdout) group. All parameters put on the same scale using data scaling method. In the learning process, parameter tuning was implemented for each method. Caret package parameter tuning was implemented using the train control function in R software, and the expand grid function supported the search process (Kuhn et al., 2020).To estimate the generalization error, 25 numbers of subsamples were created as training groups using bootstrapping resampling method. The bootstrap method is a resampling approach used to calculate statistics on a population by iteratively sampling a dataset with replacement. It is a widely applicable and powerful statistical technique used to measure the uncertainty associated with a given statistical learning method (James et al., 2013). Since we work with a small sample size in this study, the use of boot strapping is very relevant to estimation problems with small sample size and unknown distribution of the actual population (Karthik and Abhishek, 2019).

For each model, the best tuning parameters were selected using Root Mean Squared Error (RMSE) metric. Then, the seven models were combined in Caret List function using the best tuning parameters on the training group using bootstrapping resampling procedure. The Caret List function is used as it supports building lists of caret models on the same training data.

2.4. Ensemble methods

Ensemble methods are the third major machine learning approach applied. Ensemble learning is the process of learning from the advantage of combining multiple algorithms for better model performance. In applying a single method of supervised algorithms, the task is to search for a solution in parameter space, while ensembles combine multiple parameter spaces to form a better hypothesis. The above seven base models were combined in an ensemble learning called Stacking or Super Learning (Wolpert, 1992) that trains a second-level meta learner to get the optimal combination of the base learners (Laan et al., 2007).

In this study, using Caret Stack function we applied the 5 fold cross-validation resampling method on the training dataset. Cross-validation is one of the commonly used techniques for model evaluation and considered as a better technique than residual-based metrics (Kohavi, 1995). This method ensures that every data point gets to be in a test set exactly once and disadvantageous to when used in large dataset due to increasing computation cost, which was not relevant as this study applied small dataset (Karthik and Abhishek, 2019).

Outputs from the seven base models combined in a Caret Stack using each of the seven methods, which resulted in seven ensemble versions (Deane-Mayer and Knowles, 2019). The seven ensemble models include ensemble Regularized Regression (en.GLMNET), ensemble Generalized Linear Regression (en.GLM), ensemble Neural Network (en.NNET), ensemble K-nearest Neighbors (en.KNN), ensemble Recursive Partitioning and Regression Trees (en.RPART), ensemble Support Vector Machine (en.SVM) and ensemble Random Forest (en.RF).

The RMSE metric used to select the best tuning parameters. Finally, the performance of the models was measured on the holdout group. The overall steps used are presented in Fig. 2.

Finally, in addition to the RMSE metric, this study used scatter analysis to validate the performance of machine learning methods. In particular, scatter plots of measured grain versus estimated grain values prepared for training and holdout groups helped to validate model performances (Wang et al., 2019).

3. Result

3.1. Cloud restoration and prediction accuracy

We implemented gapfill on images from the study season using 21 images. The demand for high performance computing (HPC) challenged the implementation of the gapfill. Hence, we divided the study area into smaller spatial units such as tiles of size 100 pixels by 100 pixels. This enabled us to get prediction results for all images except where cloud coverage per a sub-image was 100%. Accordingly, in Fig. 3, the overall cloud percentage per the subset was 16%; all the sub-images were gapfilled except in a sub-image dated October 11 of the year 2019. To offer prediction results, gapfill requires non-zero observation values per a sub-image as evidenced on October 1, 2017, and October 16, 2017.

The detailed visual examination of prediction outputs showed the presence of differences in spatial coherence property compared to neighboring pixels. For example, sub-images with smaller cloud

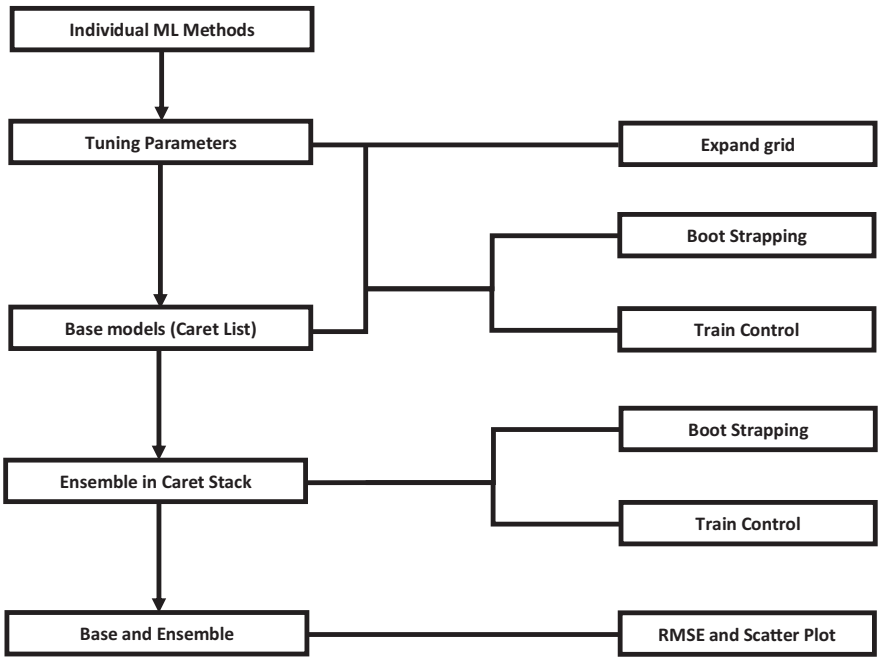


Fig. 2. Showing major learning steps applied for base and ensemble models.

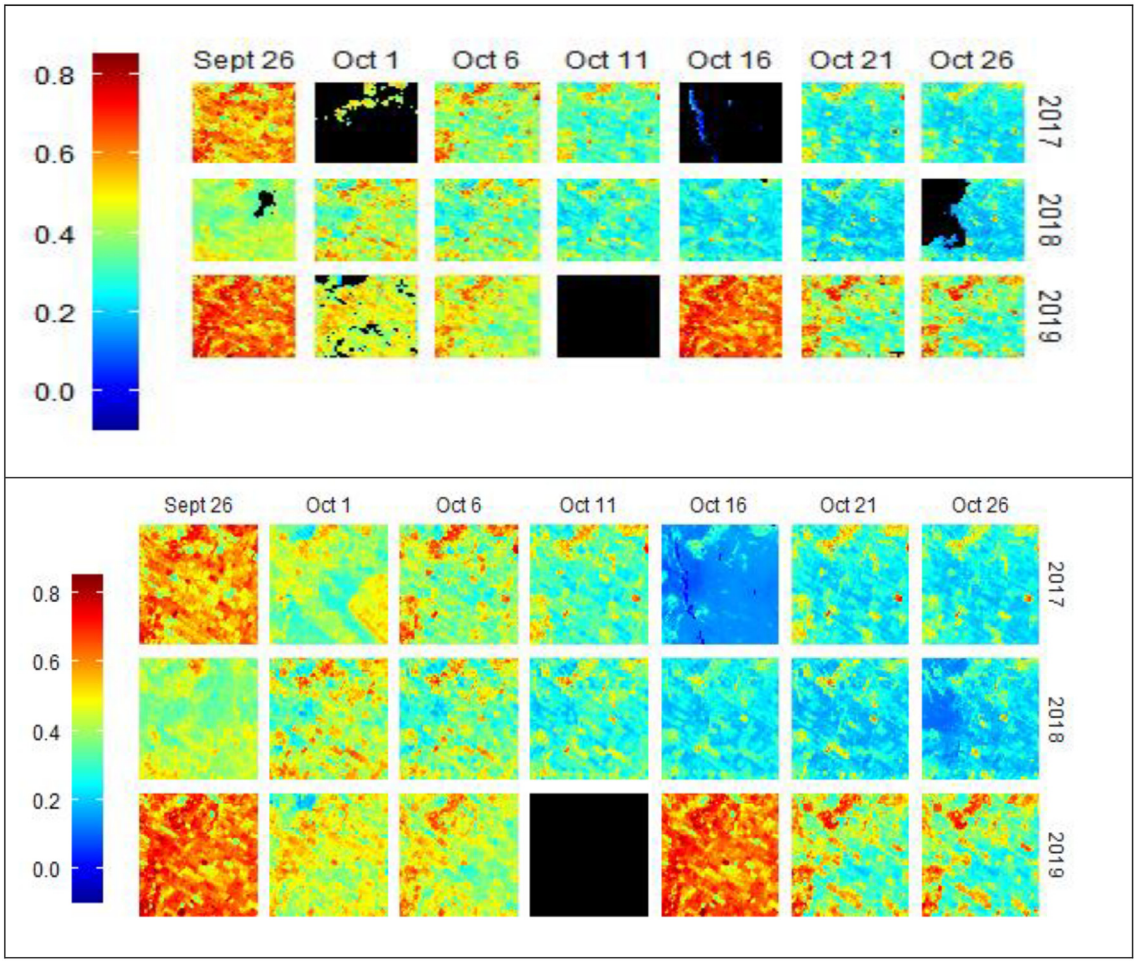


Fig. 3. A subset of 100 *100 pixels of NDVI with 16% cloud percentage (top) and the corresponding gapfilled image (bottom).

Table 1
Validation of the gapfill prediction output across a range of cloud percentage.

No	Column	Cloud % per subset	RMSE
1	1	19%	0.06
2	1	40%	0.04
3	1	52%	0.08
4	7	20%	0.06
5	7	46%	0.23
6	7	50%	0.11
7	15	28%	0.04
8	15	54%	0.07
9	15	58%	0.26

coverage such as September 26, 2018, and October 1, 2019, have a uniform texture; whereas, an image dated October 16, 2017, that has large cloud coverage shows a different and unique texture compared with the adjacent images. This indicates the cloud percentage per sub-image influences the spatial coherence and texture of the image. Besides, outputs from quantitative analysis presented in Table 1 revealed similar result.

As expected with an increasing percentage of cloud coverage per subset, the RMSE increases too. For instance, in column 15, the RMSE has increased from 0.04 to 0.07 and then to 0.26 as the cloud percentage increases from 28% to 54% and then to 58%. This is because with increasing cloud coverage the prediction method iteratively enlarges the prediction subset along the spatio-temporal dimension to get the required statistics. It could add also heterogeneous values in the subset that increases the probability of offering poor prediction estimates. In addition, it contradicts the theoretical assumption that a much narrow prediction subset is expected to offer good quality prediction in line with Tobler's first law of geography that states near things are more related than distant things.

3.2. Challenges of the restoration process

The original work of gapfill prediction method implemented on MODIS dataset that has a medium spatial resolution (500 m). In northern Alaska, gapfill was applied using 48 images that comprise missing values of $\approx 3.7 \times 10^{6}$ using 80 cores of an Intel Xeon CPU E7–2850 at 2 GHz and it took 10 h. In this study, we applied gapfill on NDVI derived from S2 with 10 m spatial resolution and 5 days temporal resolution. Such an increment in spatial resolution resulted to increase the missing values largely. In the meantime, our goal was to test the feasibility of the prediction method in an environment with limited computing capacity. In this regard, we tiled the study area along the x and y dimensions with an approximate dimension of 100*100 pixels. Accordingly, for an array of dim1 = 100, dim2 = 100, dim3 = 7, dim4 = 3, with a total number of data values of 210,000, the data restoration process took an average of 9 min using Intel(R) Core(TM) i9-9900k cpu @3.60 GHz 3.60 GHz with 32 GB Ram and with 8/16 number of Cores/Threads. For 5.7×10^{7} observation pixels, it took $\approx$ 41 h to predict. In the absence of HPC, we used tiling as a solution; nonetheless, it compromised the quality of the outputs in two ways. First, it creates lines along the boundaries of the tiles in contrast to the seamless boundary. Second, it creates sub-images with 100% cloud coverage that do not restore because of the inherent property of the gapfill algorithm. For instance, Fig. 4 demonstrates two cases where tiling creates boundary lines. On the left image where high NDVI values dominated, the tile boundaries are more visible compared to the right one where low values are prominent.

The limitation of the gapfill in sub-images with 100% cloud coverage is illustrated using Fig. 5. Sub-image dated October 11, 2019, returned no prediction values. The computing demand could be reduced using smaller tiles. However, as the size of the tile decreases, the number of sub-images with 100% cloud coverage increases. For instance, on the left side of Fig. 5 prepared using four tiles each with 100*100 dimension showed fully unrestored (NA) output on the top left corner. Whereas, the corresponding image on the right using 200*200 pixel dimension fully recovered.

Hence, tiling implementation need to consider techniques that avoid the creation of sub-images with full cloud coverage taking the tradeoff between the available computing infrastructures versus the size of the tile. In this research, as we applied square tiles we obtained some sub-images with partial restoration. Nonetheless, as shown in Fig. 6 pixels restored for most of the study area.

3.3. Data mining techniques

3.3.1. Linear regression

Under Section 3.2, phenology-based cloud restoration implemented using gapfill. The restored NDVI used to create the NDVI variable called sum-NDVI. Sum-NDVI is the sum of all NDVI values per each of the study farm boundaries (Mirasi et al, 2019). We also validated the accuracy of the restoration process. In this section as a prediction problem, we implement linear regression analysis by combining observations both from original cloud-free and restored images. The addition of observations from the restored image increases the total number of observations. Nevertheless, whether such increment positively influences the prediction problem should be explained. Hence, the purpose of the analysis is to identify how the inclusion of observations from restored images affects the regression process.

Accordingly, for images dated 2017/10/01 and 2017/10/16, the restoration process offered newly restored observations of 20 and 22 with the corresponding adj R^2 values of 0.40 and 0.54 (Table 2 fully restored column). An increase in the number of observations for the combined analysis is associated with an increase in prediction metrics. In particular, images from 2017/10/11, 2018/10/01, 2018/10/21 and

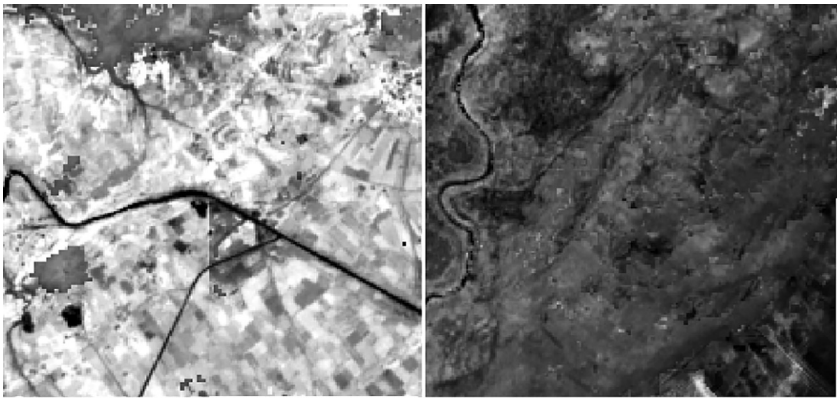


Fig. 4. Two groups of NDVI mosaicked using four tiles (each with 100*100 pixels size), on the left tile boundaries exist and on the right closer to seamless mosaic (no boundary lines).

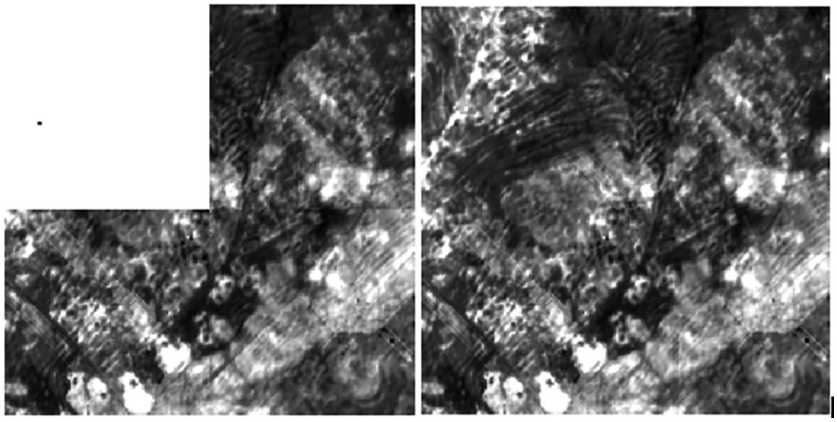


Fig. 5. Gapfilled images of NDVI prepared by mosaicking four tiles of 100*100 pixel dimension(left) and images using 200*200 pixel dimension (right).

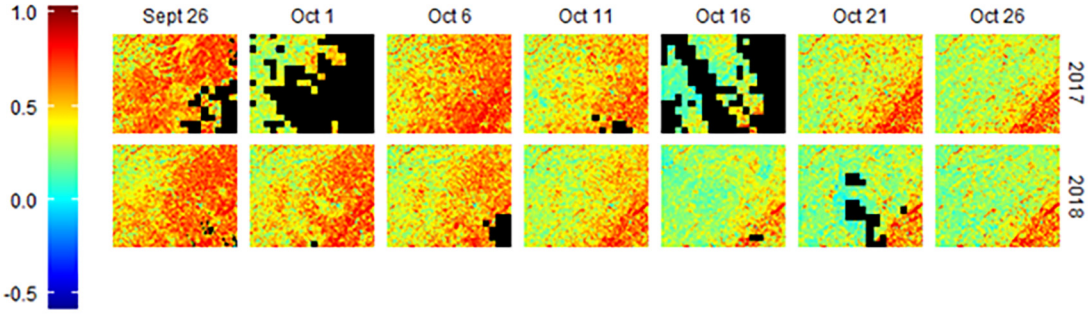


Fig. 6. Gapfilled NDVI images of the full study area for the two study years with unrestored areas depicted as black spots).

2018/10/26 that got increment of observations resulted in improvement of adj R^2 values from 0.44, 0.56, 0.27 and 0.42 to 0.55, 0.57, 0.41 and 0.46 respectively**.** This revealed the incorporation of restored observations in the regression process is advantageous in increasing prediction capability and implies the values are good enough.

Nonetheless, an image from the date of 2017/09/26 did not show any improvement. Furthermore, in 2018/10/16 image, though the total number of observations increased from 40 to 67, the adj R^2 decreased from 0.65 to 0.44. In this study, we applied similar tuning parameters across the spatio-temporal image set; nonetheless, the performance of these tuning parameters could vary across the scenes in space and time domain. This in turn could influence the quality of the restored images and ultimately the derived variables.

The advantage of restored values for a possible increment of yield prediction also assessed using multiple linear regressions (Table 3). Accordingly, the multiple linear regression output do not show improvement over univariate regression which is presented under Table 2. Besides, in light of boosting the prediction potential, we added two additional predictors, namely, NPS and Urea. The result has shown an increase in prediction potential from 0.54 using sum-NDVI composite to 0.65 using the combined dataset (sum-NDVI plus inputs). Such increment in prediction capability is resulted from the application of NPS and Urea and showed significance levels at 99.9 and 95% confidence intervals respectively (See Table 3).

3.3.2. Machine learning methods

The study applied seven machine learning methods for regression problems using three datasets: sum-NDVI (2018), sum-NDVI & inputs (2018) and sum-NDVI (2017) (Table 4).

Table 2
Comparison of original, combined (original + restored), and fully restored observations using linear regression model for the years of 2017 and 2018.

Univariate Linear Regressions for 2017 and 2018										
No	Date	Original			Combined (Original + Restored)			Fully restored		
		R^2	*p*-value	N	R^2	p-value	N	R^2	p-value	N
1	2017/09/26***	0.51	1.60e-09	52	0.50	1.02e-10	62	–	–	–
2	2017/10/01**	–	–	–	–	–	–	0.40	0.00	20
3	2017/10/11***	0.44	5.96e-09	59	0.55	4.68e-13	67	–	–	–
4	2017/10/16***	–	–	–	–	–	–	0.54	6.26e-05	22
5	2018/10/01***	0.57	1.01e-12	62	0.57	1.68e-13	66	–	–	–
6	2018/10/16***	0.66	1.51e-10	40	0.45	4.28e-10	67	–	–	–
7	2018/10/21***	0.27	2.71e-10	15	0.41	8.48e-09	63	–	–	–
8	2018/10/26***	0.42	3.52e-08	56	0.46	2.10e-10	67	–	–	–

*** Significant at 99.9%.
** Significant at 99%.

Table 3
Multiple linear regression outputs of three groups of combined (original and restored observations) datasets and inputs: sum-NDVI of 2017, sum-NDVI of 2018 and sum-NDVI of 2018 with inputs.

No	Dataset	R^2	p-value
1	2017/10/21 + 2017/10/11 + 2017/10/06 + 2017/10/26	0.54	1.141e-10
2	2018/09/26 + 2018/10/01 + 2018/10/06 + 2018/10/11 + 2018/10/16 + 2018/10/21 + 2018/10/26	0.54	8.962e-09
3	2018/09/26 + 2018/10/01 + 2018/10/06 + 2018/10/11 + 2018/10/16 + 2018/10/21 + 2018/10/26 + NPS*** + UREA*	0.65	1.155e-09

*** Significant at 99.9%.
* Significant at 95%.

Table 4
Comparison of base models for the years of 2018 and 2017 using RMSE(t/ha) on training data using Boot Strapping Method.

Models	sum-NDVI (2018)	sum-NDVI & inputs (2018)	sum-NDVI (2017)
GLMNET	0.001	0.073	0.001
SVM	0.027	0.076	0.154
KNN	0.058	0.001	0.006
NNET	0.001	0.017	0.001
RF	0.028	0.001	0.015
RPART	0.001	0.001	0.001
GLM	0.001	0.089	0.001

In general, across the seven methods and along with the three datasets RPART, GLMNET, GLM and NNET are the four models yielding the lowest RMSE values on the training dataset. All the four models resulted in RMSE values of 0.001 t/ha for sum-NDVI dataset of the year 2018 and 2017, while RPART revealed the same result on the third dataset too. Random Forest is also among the models yielding smallest RMSE values on sum-NDVI and inputs (2018) dataset. Conversely, two models such as SVM and KNN showed in relative terms lower performance across the three datasets except KNN revealed one of the lowest outputs (0.001 t/ha) on sum-NDVI & inputs(2018) dataset.

3.3.3. Ensemble models

Under Section 3.3.2, individual base models were prepared and here outputs from the seven base models used as predictors and trained to prepare seven ensemble models (Table 5).

Accordingly, two ensembles: en.GLMNET and en.RPART are the two models with better training performance. For instance, on sum-NDVI of 2018 dataset, en.GLMNET (0.012 t/ha), en.GLM (0.015 t/ha) and RPART (0.018 t/ha) produced the lowest values of RMSE. On sum-NDVI & inputs (2018) dataset, en.RPART (0.018 t/ha), en.GLMNET (0.056 t/ha), en.GLM (0.061 t/ha) and en.RF (0.051 t/ha) offer the lowest values of RMSE. On the sum-NDVI (2017) dataset, the seven ensemble versions revealed closer outputs; nonetheless, en.GLMNET and en.RPART produced some of the optimum results within the group.

3.3.4. Validation on holdout group

Section 3.3.2 presented results from base models on the training group. In this section, outputs using three datasets on the holdout group presented (Fig. 7).

Accordingly, the GLMNET, GLM and NNET models, which show better performance on the training group, revealed good generalization on sum-NDVI (2018) and sum-NDVI (2017) datasets. Regularized Regression (GLMNET) model showed a slight increase of RMSE from 0.001 t/ha on the training group to 0.006 t/ha on holdout group using sum-NDVI (2018) dataset. Likewise, the GLM model on sum-NDVI (2017) showed a slight increase of RMSE from 0.001 t/ha on the training group to 0.015 t/ha on holdout group. The neural net model has RMSE of 0.001 t/ha, 0.001 t/ha, and 0.017 t/ha on the training group compared to 0.033 t/ha, 0.006 t/ha, and 0.09 t/ha on hold out group for datasets of sum-NDVI (2018), sum-NDVI (2017) and sum-NDVI & inputs (2018) respectively.

Conversely, RPART and RF in relative terms showed overfitting across the three datasets. For instance, RPART and RF models on sum-NDVI (2018) have RMSE 0.001 t/ha and 0.028 t/ha on the training dataset that increased to 0.161 t/ha and 0.136 t/ha on the holdout group. On the other hand, SVM and KNN models demonstrate properties of underfitting where performance on the training is lower than the holdout.

In general, ensemble models expected to outperform base models. Fig. 8 presented the comparison of base models and ensembles across the three datasets on holdout group. On sum-NDVI and inputs (2018) dataset, ensemble models constitute two of the top four models i.e., GLMNET (0.001 t/ha), GLM (0.001 t/ha) en.RPART (0.015 t/ha), and en.GLMNET (0.045 t/ha). Similarly, on sum-NDVI (2018) dataset, GLMNET (0.001 t/ha), GLM (0.001 t/ha), en.GLMNET (0.015 t/ha), KNN (0.01 t/ha) and en.RPART (0.015 t/ha) are the top five models of which the two are ensembles. On the other hand, a comparison using sum-NDVI (2017) dataset showed that an ensemble model, en.RF (0.008 t/ha), is among the top four models. Nonetheless, ensembles such as en.SVM (0.245 t/ha) and en.NNET (0.172 t/ha) are among the least performing ones on sum-NDVI (2017) dataset.

Table 5
Seven ensemble models using the seven base models as input on training data (RMSE are reported in t/ha).

Models	sum-NDVI (2018)	sum-NDVI & inputs (2018)	sum-NDVI (2017)
en.GLMNET	0.012	0.056	0.163
en.SVM	0.099	0.165	0.311
en.KNN	0.298	0.109	0.079
en.NNET	0.354	0.180	0.194
en.RF	0.051	0.123	0.166
en.RPART	0.018	0.018	0.128
en.GLM	0.015	0.061	0.166

3.3.5. Model validation using scatter plot analysis

Model comparisons presented under Sections of 3.3.2, 3.3.3 and 3.3.4 applied RMSE measure. To further consolidate and verify performances in this part graphical analysis is used. In particular, scatter plots for measured grain yield versus estimated grain yield implemented for models that revealed better performance based on RMSE metric. A perfect scatter plot demonstrated two characteristics. First, the 1:1 diagonal line should pass through the origin of the plot. Second, the range of estimated values (y-axis) should be equal to the range of measured values (x-axis).

Fig. 9 presented six scatter plots for the sum-NDVI (2018) dataset for three base models where each model has two plots for the training group and its holdout equivalent. Thus, based on scatter plot analysis, GLM, RF and GLMNET in their order are the three better models. In particular, the GLM model revealed better performance both on the training and holdout group because the 1:1 diagonal line placed closely to the origin and the good match between the range of values along the x and y axes. Besides, the performance of these models also verified using RMSE error presented under Section 3.3.4.

Based on RMSE measure ensemble models such as en.GLMNET and en.GLM were also among the best models as discussed under Section 3.3.4. Nonetheless, their scatter plot performance does not support the RMSE performance. Fig. 10 presented eight scatter

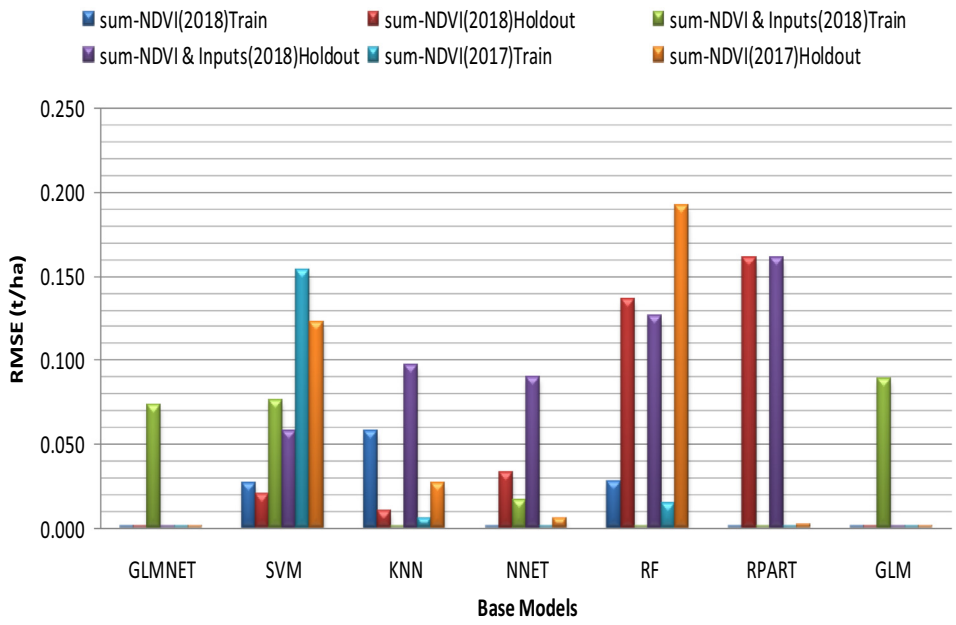
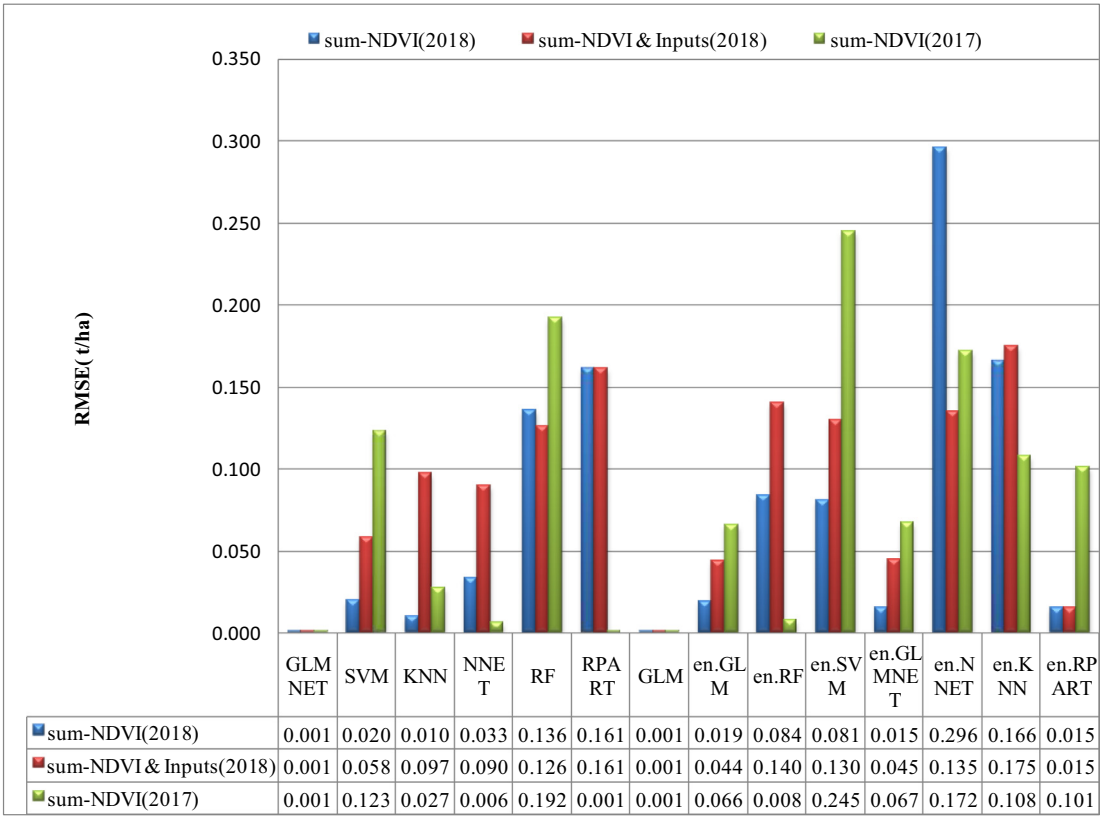


Fig. 7. Bar graph showing performance of base machine learning methods across three datasets using RMSE (t/ha).

plots for four better ensemble models. Accordingly, ensembles models showed poor performance, i.e., lower generalization on the holdout group. To be exact, the diagonal lines showed a clear shift from the origin on the holdout group. Consequently, based on RMSE and graphical measures ensemble models do not show improvement over base models.

3.3.6. Ensembles based on ranking of base learners

The ensemble models, presented under Section 3.3.5, were implemented using seven base learners as input and do not show improvement over base models. Nonetheless, previous studies revealed, to develop an ensemble learner with better predictive potential, two conditions need to be met (Krogh and Vedelsby, 1995; Zhou, 2009). First,



	GLM NET	SVM	KNN	NNE T	RF	RPA RT	GLM	en.GL M	en.RF	en.SV M	en.GL MNE T	en.N NET	en.K NN	en.RP ART
sum-NDVI(2018)	0.001	0.020	0.010	0.033	0.136	0.161	0.001	0.019	0.084	0.081	0.015	0.296	0.166	0.015
sum-NDVI & Inputs(2018)	0.001	0.058	0.097	0.090	0.126	0.161	0.001	0.044	0.140	0.130	0.045	0.135	0.175	0.015
sum-NDVI(2017)	0.001	0.123	0.027	0.006	0.192	0.001	0.001	0.066	0.008	0.245	0.067	0.172	0.108	0.101

Fig. 8. Comparison of base models and ensembles on holdout group in RMSE (t/ha) for three dataset groups.

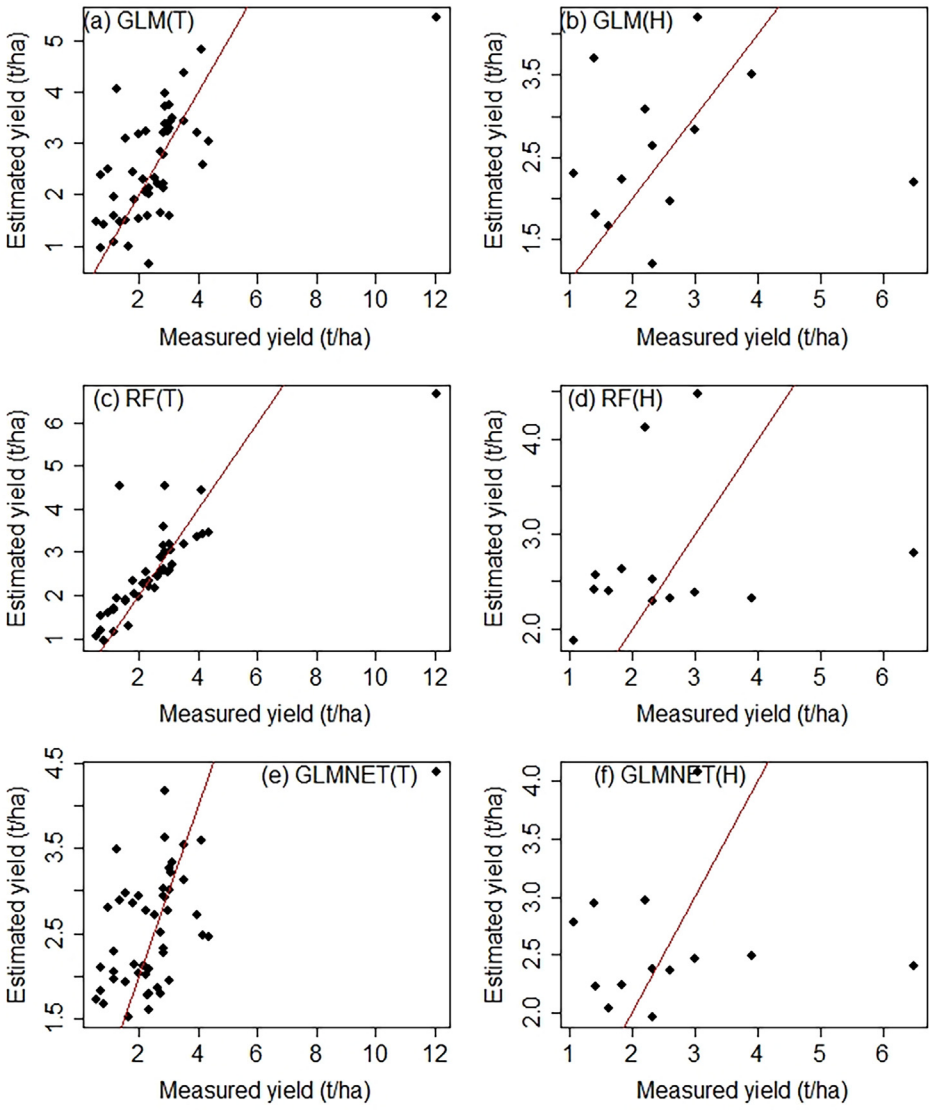


Fig. 9. Scatter plot showing graphical performance of better performing base models for sum-NDVI(2018) dataset. On the left column, T represents performance on training group; on the right column, H represents performance on holdout group. The diagonal lines showed the 1:1 relationship.

the input base learners should have sufficient diversity. Second, each of the individual base learners ought to be as accurate as possible. Practically, methods such as sub-sampling the training examples, modifying the attributes, manipulating the outputs, adding randomness into learning methods, or even using multiple ways at the same time help to achieve the diversity of the base learners (Krogh and Vedelsby, 1995; Zhou, 2009). The accuracy of learners could be estimated using methods such as cross-validation and holdout. On the other hand, Yang and Lv, 2021 described three groups of strategies for selecting the base learner such as clustering-based methods, ranking-based methods and selection-based methods. The ranking-based method first applies a certain accuracy measurement to rank the base learners. Then, it uses a suitable stopping criterion to choose some base learners for creating an ensemble (Martinez-Munoz et al., 2008).

Therefore, in this study, we adopted the ranking-based method using the performance of base learners on the holdout group. That is, ensembles are developed using outputs from the three best performing base learners such as GLM, RF and GLMNET models as predictors. Once again, the seven machine learning methods used to create seven ensemble versions. Thus, Fig. 11 revealed scatter plots for the top four ensembles based on ranking method.

Accordingly, the same type of ensembles such as en.RF, en.KNN, en.GLM and en.GLMNET that were better using the full base learners found to be also the better ones using the ranking method too. Nonetheless, the new sets of ensembles based on ranking method do not show improvement over the first group of ensembles using full base learners. In summary, in this study, though ensembles developed using two different approaches, they did not result in improved performance over base learners. This could be associated with the limited diversity of base learners and the use of a small dataset where each observation point has a large weight on model performance affecting model stability.

4. Discussion

We applied gapfill algorithm on NDVI dataset derived from S2 images having a high spatial and temporal resolution. The pioneer research using gapfill used MODIS NDVI product with eight days of temporal resolution and 500 m spatial resolution in Northern Alaska. The study used images from six years period: 2004–2009 and from the dates of May 24 to September 13. In such a setup, the good data per day of the year was 30%, and 48 images (Gerber et al., 2018).

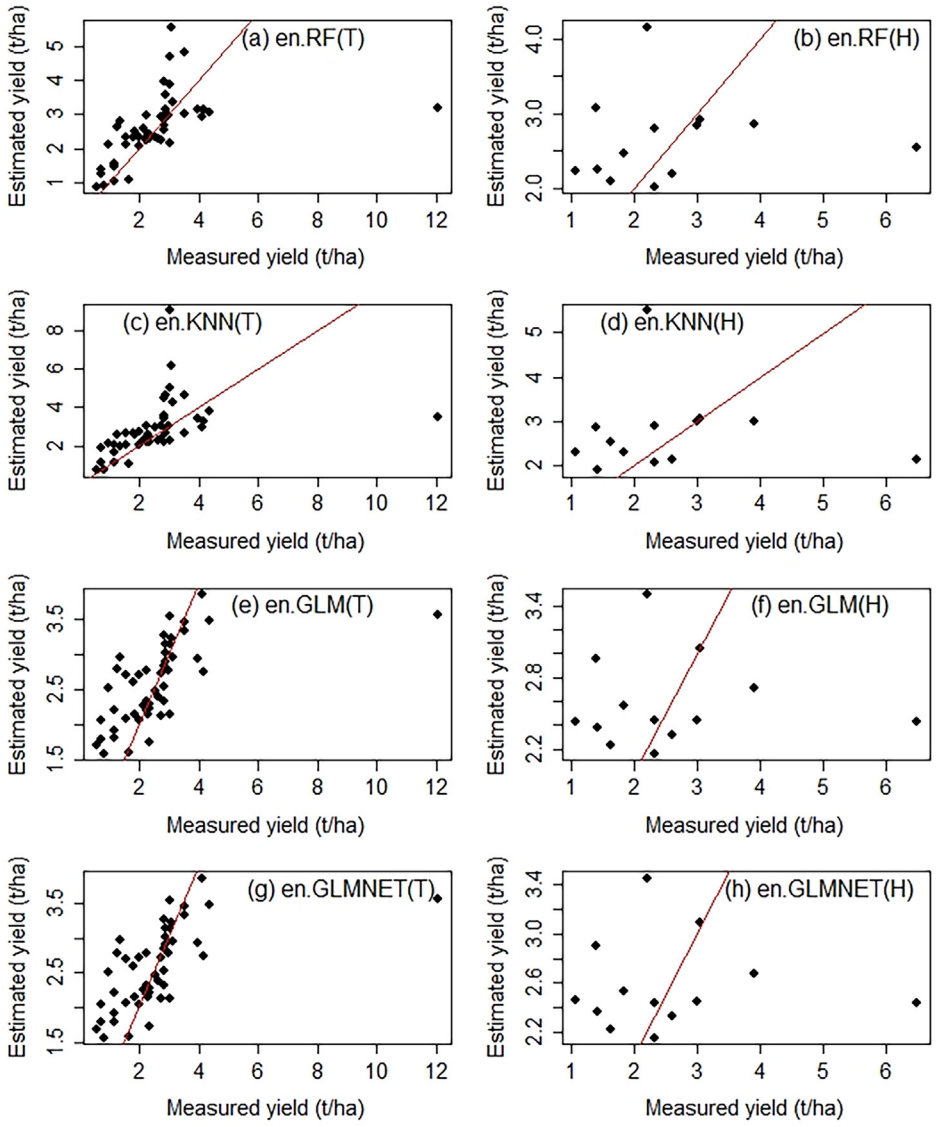


Fig. 10. Scatter plot showing graphical performance of better performing four ensemble models for sum-NDVI(2018) dataset. On the left column, T represents performance on the training group; on the right column, H represents performance on the holdout group. The diagonal lines showed the 1:1 relationship.

In this study, we used images from three years period (2017-2019) and from August 26 to September 26 making 21 images. The tiling process affected the percentage of cloud coverage in sub-images. It created fully covered sub-images; hence, the good data per day of the year was as low as 0%. Consequently, the gapfill did not result in a full restoration of the study area. In contrast, in the original study as the maximum cloud percentage was 70% all the missing values restored. Yet, on both of the studies, the prediction accuracy decreases with the increment of cloud percentage per subset. In the pioneer study, for cloud percentages of 20%, 30%, 40% and 50% the RMSPE (Root Mean Square Percentage Error) were 41.86, 42.54, 41.34, and 59.58 respectively (Gerber et al., 2018).

Therefore, in terms of getting a fully restored subset, the cloud percentage per day of the year is the most important parameter. In this study, we used square tiles that resulted in full cloud cover in some sub-images and get predictions for most of the study area. Future studies might consider other tiling procedures that potentially avoid or minimizes the number of sub-images with full cloud cover.

In this study, the gapfill algorithm implemented using the default configuration. It begins by identifying a four-dimensional array around the missing values to predict. Then, the subset has $\lambda x = \lambda y = 5$ pixels along the x and y dimensions from the missing value, $\lambda s = 1$step in both directions of the seasonal index and, $\lambda a = 3$ years in both directions of the year index. Thus, the major difference between the current study and the pioneer study is on the value of the year index. In the pioneer study (Gerber et al., 2018), $\lambda a = 6$ years was used, while the current study used $\lambda a = 3$. The two studies have closer values of the maximum seasonal index, i.e., on the pioneer research λs could have a maximum value of 8 while the current study could have a maximum $\lambda s = 7$.

In applying gapfill, in this study, we started by considering two key challenges. First, since the S2 sensor is recently available, images are available only for three years. Second, **S2** is a high spatial resolution sensor (10 m) making the total number of missing values to be too large. Hence, it demands the use of high computing infrastructure, which is unfortunately not available in many situations.

Given such limitations, the study has successfully applied gapfill under the current study area. In this study, our main goal was to understand the predictive skill of the sum-NDVI parameter using original and restored observations using gapfill. Except for a single date (2018/10/16), the incorporation of restored observations in regression analysis offered an increased prediction skill. Thus, the gapfill method offered

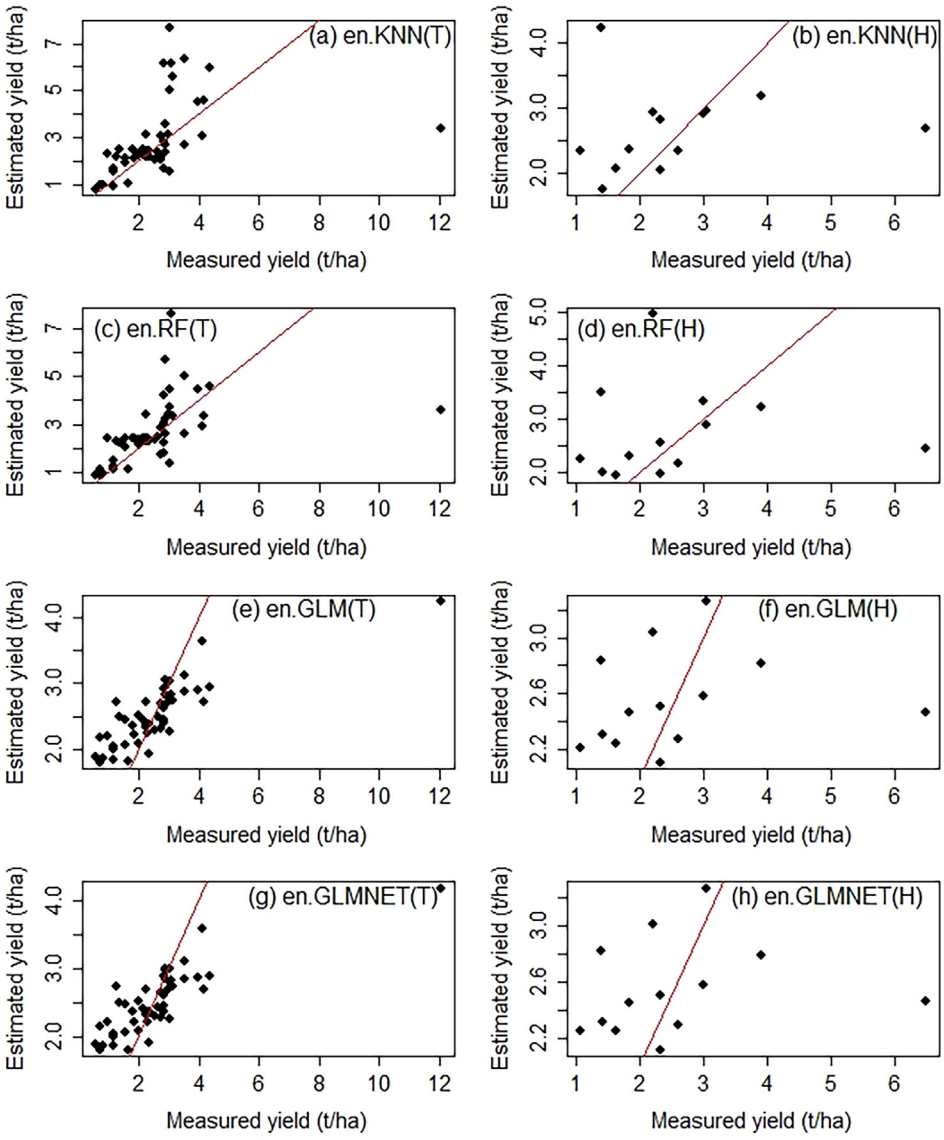


Fig. 11. Scatter plot showing graphical performance of better performing four ensemble models based on ranking method for sum-NDVI (2018) dataset. On the left column, T represents performance on the training group, on the right column; H represents performance on the holdout group. The diagonal lines showed the 1:1 relationship.

quality estimates for cloudy pixels that are useful in empirical-based yield estimation with positive influence in the final models.

In this study, ensemble models do not show improvement over base models. Base models such as GLM, RF and GLMNET revealed RMSE of 0.001 t/ha, 0.136 t/ha and 0.001 t/ha for sum-NDVI(2018) dataset as well as 0.001 t/ha, 0.192 t/ha and 0.001 t/ha for sum-NDVI(2017) dataset. In contrast, RPART and RF models in relative terms showed overfitting across the three datasets. On the other hand, models such as SVM and KNN demonstrate properties of underfitting in which performance on the training is lower than the holdout. The addition of other hyperparameters besides the limited parameters used in the current study could improve the performance of these models.

The superior performance of the GLM algorithm could be associated with the inherent nature of the model and the relationship between the response and predictor variables. Generalized Linear Model used to handle non-linear relationship as it transforms the original response variable using the link function to be linearly related to the independent variables (Hardin et al., 2007). As previous studies indicated, the majority of the relationship between wheat yield and sum-NDVI is linear (Kumhálová and Matějková, 2017). Yet, there is some non-linearity (Zhao and Cen, 2013; Ram, 2021) that needs to be explained. Hence, the use of the GLM with Gaussian distribution and identify link function will transform such non-linearity to linear model offering increased model representation.

Random Forest is an ensemble algorithm that use bootstrap aggregation method, in particular, it computes average prediction from various decision trees predictions. This enables RF models to be less influenced by outliers. This study is implemented using small number of observation and decided to keep all the sample observations that potentially constitute some outliers. Thus, algorithms such as RF, which are less prone to outliers, are expected to yield increased performance. Besides, RF algorithm is powerful in handling both linear and non-linear relationship (Murthy, 2020; Trehan, 2020).

The ability of machine learning to learn information directly from observed data is commonly applied using big dataset, whereas ML for small dataset is a new and growing area of research and development (Forman and Cohen, 2004; Shaikhina et al., 2014). This is associated partly with, in some applications, for instance, in biomedical and material sciences; the collection of samples requires high cost (Feng et al., 2019). The training and initialization routines of machine learning algorithms involve randomization that commonly improves the convergence of the learning process to the global minimum. Nonetheless,

this process negatively affects the stability and generalization capacity of algorithms. In using small dataset, the problem becomes more pronounced (Forman and Cohen, 2004).

As a growing field of research, the use of ML under small dataset domain various across various disciplines. In some fields, for instance, in materials physics and chemistry, numerous studies applied ML to small dataset. A backpropagated Neural Network model using 53 points used successfully to predict the properties of ultrahigh-performance concrete (Ghafari et al., 2015). A strategy named as crude estimation property was proposed to improve ML on small datasets (around 100 data points) (Zhang and Ling, 2018). Similarly, Shaikhina et al., 2015 designed neural net and decision tree algorithms for prediction of antibody-mediated kidney transplant rejection using 35 bone specimens and 80 kidney transplants and achieved high accuracy of 98.3% and 85% respectively.

In some previous studies, ensemble models showed superiority to base models. For example, a stacking ensemble of ANN revealed a relative RMSE of 6.8% and R^2 of 0.68 at the predicted yield in sugar cane crop using MODIS NDVI time series (Fernandes et al., 2017). The superior performance of ensemble models is also reported on other biophysical datasets. An Ensemble of gradient boosting, multi-narrative adaptive regression spline, random forest, and Support Vector Machine outperformed the individual models for surface soil organic carbon stocks (Mishra et al., 2020). An average-ensemble model is recommended as the best model in materials design using 25 numbers of observations (Vanpoucke et al., 2020).

Nonetheless, in this study, although two ensemble approaches are applied, the performances of ensemble models do not show improvement over base models.

Overall, comparison of the current outputs with previous study outputs could be problematic due to factors such as discipline difference and difference in study design and procedure. Even for same discipline, the implementation of different study design (resampling procedures, outlier treatments, data split, and the number and type of hyperparameters searched) might contribute to reveal different result. Therefore, the difference between the two models (GLM and RF) and the rest five models could be explained partly by the study design implemented.

5. Conclusion and recommendations

In this study, the application of gapfill on high-resolution imagery, S2, resulted in good quality estimates of cloudy pixels. Consequently, regression-based crop prediction methods benefited. The rare availability of HPC, especially in resource-limited environments is the major challenge of applying gapfill in high-resolution sensors. The study showed that the use of images from a critical window of the crop phenology and implementation of square tiles are the possible solutions for decreasing the demand for HPC.

The study evaluated base machine learning and ensemble methods for wheat yield prediction. Ensemble methods produced a similar performance with that of base models. Ensemble models using the seven predictors from base models as well as using three selected predictors based on accuracy following the ranking methodology showed similar performance. The study, employing sum-NDVI dataset, predicts wheat grain yield with RMSE of 0.001 t/ha and 0.136 t/ha using GLM and RF models respectively. The two models showed consistent and better performance across the three dataset groups. Besides, they showed good generalization on holdout dataset that is also supported using scatter plot analysis.

Given the small number of observation samples, the study successfully exploited the synergy of combining the spatio-temporal gapfill method, phenology and machine learning methods and revealed a reliable yield prediction approach.

Overall, satellite-based farm-level yield prediction methods are rarely available in many crop production systems. Therefore, the methods revealed in this study will be crucial in numerous projects such as food security, crop insurance among others where farm-level yield data are key inputs. Furthermore, though the study revealed yield prediction methodology for wheat, the same approach could be adopted to develop reliable remote sensing based yield prediction methods for other cereal crops including barley, maize, and sorghum.

Funding

The authors would like to thank the Ethiopian Space Science and Technology Institute for providing financial support. Besides, we acknowledge the Ethiopian Civil Service University for sponsoring the principal investigator.

Data availability statement

We would like to confirm that the data supporting the current study is available based on reasonable request.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

Arsi-Sire Wereda administration facilitated the field data collection. We acknowledge to Ethiopian Geospatial Institute and in particular to Tulu Besha (Ph.D) for offering computing infrastructure. We offer our special gratitude to Dagmawi Teklu (Ph.D) for commenting the manuscript.

References

Adams, M.L., Cook, S.E., Caccetta, P.A., Pringle, M.J., 1999, January. Machine learning methods in site-specific management research: An Australian case study. Proceedings of the Fourth International Conference on Precision Agriculture. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI, USA, pp. 1321–1333.

Aha, D.W., Kibler, D., Albert, M.K., 1991. Instance-based learning algorithms. Mach. Learn. 6 (1), 37–66.

Appelhans, T., Mwangomo, E., Hardy, D.R., Hemp, A., Nauss, T., 2015. Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania. Spatial Stat. 14, 91–113.

Atkinson, P.M., Jeganathan, C., Dash, J., Atzberger, C., 2012. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. Remote Sens. Environ. 123, 400–417.

Bishop, C.M, 1995. *Neural networks for pattern recognition*. Oxford university press.

Brasesco, F., Asgedom, D., Casari, G., 2019. Strategic Analysis and Intervention Plan for Fresh and Industrial Tomato in the Agro-Commodities Procurement Zone of the Pilot Integrated Agro-Industrial Park in Central-Eastern Oromia, Ethiopia. FAO, Addis Ababa.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., 2002. Manual On Setting Up, Using, And Understanding Random Forests. Berkeley. https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 2017. Classification and Regression Trees. Routledge.

Brownlee, J., 2016. Master Machine Learning Algorithms: Discover how they Work and Implement them from Scratch. Machine Learning Mastery.

Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., Peng, B., 2019. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. Agric. For. Meteorol. 274, 144–159.

Deane-Mayer, Z.A., Knowles, J.E., 2019. caretEnsemble: Ensembles of Caret Models. p. 35 *R package version*, *2*(1).

Dodds, F., Bartram, J., 2016. The Water, Food, Energy and Climate Nexus: Challenges and an Agenda for Action. Routledge.

Drummond, S.T., Sudduth, K.A., Birrell, S.J., 1995. Analysis and Correlation Methods for Spatial Data. ASAE Paper No. 951335ASAE, St. Joseph, Mich.

Drummond, S.T., Sudduth, K.A., Joshi, A., Birrell, S.J., Kitchen, N.R., 2003. Statistical and neural methods for site–specific yield prediction. Trans. ASAE 46 (1), 5.

Feng, S., Zhou, H., Dong, H., 2019. Using deep neural network with small dataset to predict material defects. Mater. Des. 162, 300–310.

Fernandes, J.L., Ebecken, N.F.F., Esquerdo, J.C.D.M., 2017. Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble. Int. J. Remote Sens. 38 (16), 4631–4644.

Fischer, R.A., 2011. Wheat physiology: a review of recent developments. Crop Pasture Sci. 62 (2), 95–114.
Forman, G., Cohen, I., 2004, September. Learning from little: comparison of classifiers given little training. European Conference on Principles of Data Mining and Knowledge Discovery. Springer, Berlin, Heidelberg, pp. 161–172.
Friedman, J., Hastie, T., Tibshirani, R., 2009. Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. R package version 1.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1.
Fritsch, Stefan, Guenther, Frauke, Wright, Marvin N., 2019. neuralnet: Training of Neural etworks. R package version 1.44.2. https://CRAN.R-project.org/package=neuralnet.
Garonna, I., de Jong, R., Schaepman, M.E., 2016. Variability and evolution of global land surface phenology over the past three decades (1982–2012). Glob. Chang. Biol. 22 (4), 1456–1468.
Gerber, F., 2018. Fill Missing Values in Satellite Data. Package 'gapfill'. https://git.math.uzh.ch/florian.gerber/gapfill.
Gerber, F., de Jong, R., Schaepman, M.E., Schaepman-Strub, G., Furrer, R., 2018. Predicting missing values in spatio-temporal remote sensing data. IEEE Trans. Geosci. Remote Sens. 56 (5), 2841–2853.
Ghafari, E., Bandarabadi, M., Costa, H., Júlio, E., 2015. Prediction of fresh and hardened state properties of UHPC: comparative study of statistical mixture design and an artificial neural network model. J. Mater. Civ. Eng. 27 (11), 04015017.
Gomez, D., Salvador, P., Sanz, J., Casanova, J.L., 2019. Potato yield prediction using machine learning techniques and sentinel 2 data. Remote Sens. 2019 (11), 1745.
Gunn, S.R., 1998. Support vector machines for classification and regression. ISIS Tech. Rep. 14 (1), 5–16.
Hardin, J.W., Hardin, J.W., Hilbe, J.M., Hilbe, J., 2007. Generalized Linear Models and Extensions. Stata press.
He, L., Mostovoy, G., 2019. Cotton yield estimate using Sentinel-2 data and an ecosystem model over the southern US. Remote Sens. 2019 (11), 2000.
Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B., 1998. Support vector machines. IEEE Intell. Syst. Appl. 1998 (13), 18–28.
Hird, J.N., McDermid, G.J., 2009. Noise reduction of NDVI time series: an empirical comparison of selected techniques. Remote Sens. Environ. 113 (1), 248–258.
Jaikla, R., Auephanwiriyakul, S., Jintrawet, A., 2008. Rice yield prediction using a Support Vector Regression method. Conference: Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on. Vol. 1. https://doi.org/10.1109/ECTICON.2008.4600365.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Vol. 112. springer, New York, p. 18.
Jiang, D., Yang, X., Clinton, N., Wang, N., 2004. An artificial neural network model for estimating crop yields using remotely sensed information. Int. J. Remote Sens. 25 (9), 1723–1732.
Kalluri, S.N.V., Zhang, Z., JaJa, J., Liang, S., Townshend, J.R.G., 2001. Characterizing land surface anisotropy from AVHRR data at a global scale using high performance computing. Int. J. Remote Sens. 22 (11), 2171–2191.
Karthik, R., Abhishek, S., 2019. Machine Learning Using R: With Time Series and Industry-Based Use Cases in R.
Kbakural, B.R., Robert, P.C., Huggins, D.R., 1999, January. Variability of corn/soybean yield and soil/landscape properties across a southwestern Minnesota landscape. Proceedings of the Fourth International Conference on Precision Agriculture. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI, USA, pp. 573–579.
Kim, N., Ha, K.J., Park, N.W., Cho, J., Hong, S., Lee, Y.W., 2019. A comparison between major artificial intelligence models for crop yield prediction: case study of the midwestern united states, 2006–2015. ISPRS Int. J. Geo Inf. 8 (5), 240.
Kitchen, N.R., Sudduth, K.A., Drummond, S.T., 1999. Soil electrical conductivity as a crop productivity measure for claypan soils. J. Prod. Agric. 12 (4), 607–617.
Kohavi, R., 1995, August. A study of cross-validation and bootstrap for accuracy estimation and model selection. Ijcai 14 (2), 1137–1145.
Krogh, A., Vedelsby, J., 1995. Neural network ensembles, cross validation, and active learning. Adv. Neural Inf. Proces. Syst. 7, 231–238.
Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., 2020. caret: Classification and Regression Training. R package version 6.0–86. Avaliable at: https://cran.r-project.org/web/packages/caret/caret.Pdf.
Kumhálová, J., Matějková, Š., 2017. Yield variability prediction by remote sensing sensors with different spatial resolution. Int. Agrophys. 31 (2), 195.
van der Laan, M.J., Polley, Eric C., Hubbard, Alan E., 2007. Super learner. Statistical Applications in Genetics and Molecular Biology. 6.
Lee, Craig, Gasster, Samuel, Plaza, Antonio, Chang, Chein-I, Huang, Bormin, 2011. Recent developments in high performance computing for remote sensing: a review. selected topics in applied earth observations and remote sensing. IEEE J. 4, 508–527. https://doi.org/10.1109/JSTARS.2011.2162643.
Lobell, D.B., Thau, D., Seifert, C., Engle, E., Little, B., 2015. A scalable satellite-based crop yield mapper. Remote Sens. Environ. 164, 324–333.
Martinez-Munoz, G., Hernández-Lobato, D., Suárez, A., 2008. An analysis of ensemble pruning techniques based on ordered aggregation. IEEE Trans. Pattern Anal. Mach. Intell. 31 (2), 245–259.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. 2nd ed. Chapman & Hall, London.
Meessen, J.H., Petersen, H., 2010. Urea. Ullmann's Encyclopedia of Industrial Chemistry. Wiley-VCH, Weinheim https://doi.org/10.1002/14356007.a27_333.
Mirasi, A., Mahmoudi, A., Navid, H., Valizadeh Kamran, K., Asoodar, M.A., 2019. Evaluation of sum-NDVI values to estimate wheat grain yields using multi-temporal Landsat OLI data. Geocarto Int. 1–16.
Mishra, U., Gautam, S., Riley, W., Hoffman, F.M., 2020. Ensemble machine learning approach improves predicted spatial variation of surface soil organic carbon stocks in data-limited northern circumpolar region. Front. Big Data 3, 40.
Murthy, S., 2020. Be the Outlier: How to Ace Data Science Interviews. New Degree Press, Kindle Edition.
Nelder, J.A., Wedderburn, R.W., 1972. Generalized linear models. J. Royal Stat. Soc. Series A (General) 135 (3), 370–384.
Pektürk, M.K., Ünal, M., 2018, August. Performance-aware high-performance computing for remote sensing big data analytics. Data Mining. BoD–Books on Demand, p. 69.
Piccini, Nathan, 2019. 101 Machine Learning Algorithms for Data Science with Cheat Sheets. datasciencedojo(blog). August 21 https://online.datasciencedojo.com/blogs/101-machine-learning-algorithms-for-data-science-with-cheat-sheets.
Potgieter, A.B., Power, B., Mclean, J., Davis, P., Rodriguez, D., 2014. Spatial estimation of wheat yields from Landsat's visible, near infrared and thermal reflectance bands. Int. J. Remote Sens. Appl 4, 134–143.
Ram, P., Apr 27, 2021. "Generalized Linear Models | What does it mean?" Great Learning Team (blog). https://www.mygreatlearning.com/blog/generalized-linear-models/.
RSrivastava, T., 2020. Introduction to k-Nearest Neighbors: A powerful Machine Learning Algorithm (with implementation in Python & R). Analytics Vidhya, Mar. 26, 2018.
Shaikhina, T., Khovanova, N.A., Mallick, K.K., 2014, June. Artificial neural networks in hard tissue engineering: another look at age-dependence of trabecular bone properties in osteoarthritis. IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, pp. 622–625.
Terry, T., Beth, A., 2019. rpart: Recursive Partitioning and Regression Trees. R package version 4.1–15. https://CRAN.R-project.org/package=rpart.
Therneau, T.M., Atkinson, E.J., 2019. Mayo Foundation (An introduction to recursive partitioning using the RPART routines).
Trehan, D., July 2, 2020. "Why Choose Random Forest and Not Decision Trees". TOWARDS AI (blog). https://towardsai.net/p/machine-learning/why-choose-random-forest-and-not-decision-trees.
Van Den Bergh, F., Wessels, K.J., Miteff, S., Van Zyl, T.L., Gazendam, A.D., Bachoo, A.K., 2012. HiTempo: a platform for time-series analysis of remote-sensing satellite data in a high-performance computing environment. Int. J. Remote Sens. 33 (15), 4720–4740.
Vanpoucke, D.E., van Knippenberg, O.S., Hermans, K., Bernaerts, K.V., Mehrkanoon, S., 2020. Small data materials design with machine learning: when the average model knows best. J. Appl. Phys. 128 (5), 054901.
Verger, A., Baret, F., Weiss, M., Kandasamy, S., Vermote, E., 2013. The CACAO method for smoothing, gap filling, and characterizing seasonal anomalies in satellite time series. IEEE Trans. Geosci. Remote Sens. 51 (4), 1963–1972.
Wang, J., Dai, Q., Shang, J., Jin, X., Sun, Q., Zhou, G., Dai, Q., 2019. Field-scale rice yield estimation using sentinel-1A synthetic aperture radar (SAR) data in coastal saline region of Jiangsu Province, China. Remote Sens. 11 (19), 2274.
Weiss, D.J., Atkinson, P.M., Bhatt, S., Mappin, B., Hay, S.I., Gething, P.W., 2014. An effective approach for gap-filling continental scale remotely sensed time-series. ISPRS J. Photogramm. Remote Sens. 98, 106–118.
Wendroth, O., Jürschik, P., Nielsen, D.R., 1999. Spatial crop yield prediction from soil and land surface state variables using an autoregressive state-space approach. Precision agriculture'99, Part 1 and Part 2. Papers presented at the 2nd European Conference on Precision Agriculture, Odense, Denmark, 11-15 July 1999. Sheffield Academic Press, pp. 419–428.
Wolpert, D.H., 1992. Stacked generalization. Neural Netw. 5 (2), 241–259.
Yang, Y., Lv, H., 2021. Discussion of Ensemble Learning under the Era of Deep Learning. arXiv preprint *arXiv:2101.08387*.
Zhang, Y., Ling, C., 2018. A strategy to apply machine learning to small datasets in materials science. Npj Comp. Mater. 4 (1), 1–8.
Zhao, Y., Cen, Y., 2013. Data Mining Applications with R. Academic Press.
Zhao, D., Reddy, K.R., Kakani, V.G., Read, J.J., Koti, S., 2007. Canopy reflectance in cotton for growth assessment and lint yield prediction. Eur. J. Agron. 26 (3), 335–344.
Zhao, Y., Potgieter, A.B., Zhang, M., Wu, B., Hammer, G.L., 2020. Predicting wheat yield at the field scale by combining high-resolution Sentinel-2 satellite imagery and crop modelling. Remote Sens. 12 (6), 1024.
Zhong, Y., Fang, J., Zhao, X., 2013, July. VegaIndexer: a distributed composite index scheme for big spatio-temporal sensor data on cloud. 2013 IEEE International Geoscience and Remote Sensing Symposium-IGARSS. IEEE, pp. 1713–1716.
Zhou, Z.H., 2009. Ensemble Learning, Encyclopedia of Biometrics.
Zscheischler, J., Mahecha, M.D., 2014. An extended approach for spatiotemporal gapfilling: dealing with large and systematic gaps in geoscientific datasets. Nonlinear Process. Geophys. 21 (1), 203–215.