ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Testing spatial out-of-sample area of influence for grain forecasting models

To cite this article: F Davenport *et al* 2024 *Environ. Res. Lett.* **19** 114079

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
8 April 2024

REVISED
16 September 2024

ACCEPTED FOR PUBLICATION
8 October 2024

PUBLISHED
18 October 2024

LETTER

Testing spatial out-of-sample area of influence for grain forecasting models

F Davenport[1,*], D Lee[1,2], S Shukla[1], G Husak[1], C Funk[1], M Budde[3] and J Rowland[3]

[1] Climate Hazards Center, Department of Geography, University of California, Santa Barbara, CA, United States of America
[2] Department of Civil Engineering, University of Manitoba, Winnipeg, Manitoba, Canada
[3] U.S. Geological Survey, Earth Resources Observation and Science Center, Sioux Falls, SD, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: frank_davenport@ucsb.edu

Keywords: forecasting, spatial transfer, food security, machine learning

Supplementary material for this article is available online

Abstract

We examine the factors that determine if a grain forecasting model fit to one region can be transferred to another region. Prior research has proposed examining the area of applicability (AoA) of a model based on structurally similar characteristics in the Earth Observation predictors and weights based on the model derived feature importance. We expand on and evaluate this approach in the context of grain yield forecasting in Sub-Saharan Africa. Specifically, we evaluate an AoA methodology established for generating raster surfaces and apply it to vector supported grain data. We fit a series of ensemble tree models both within single countries and across multiple sets of countries and then test those models in countries excluded from the training set. We then calculate and decompose AoA measures and examine several different performance metrics. We find that the spatial transfer accuracy does not vary across season but does vary by average rainfall and across high, medium, and low yielding regions. In general, areas with higher yields and medium to high average rainfall tend to have higher accuracy for both model training and transfer. Finally, we find that fitting models with multiple countries provides more accurate out-of-sample estimates when compared to models fitted to a single country.

1. Introduction

The potential for predictive models based on Earth observations (EO) and survey data to assist in famine early warning, agricultural outlooks, and other development applications is rapidly growing (Becker-Reshef *et al* 2010, Davenport *et al* 2021, FAO 2016, Johnson 2014, Kouadio *et al* 2014, Krell *et al* 2019, Lee *et al* 2024b, Newlands *et al* 2014, Schauberger *et al* 2020, Shukla *et al* 2021, Zhang *et al* 2019). Although the spatial-temporal extent of EO data is expansive, high quality agricultural survey data and official statistics are generally limited in spatial and temporal scope. Limitations on observed data, used to both train and validate models, raise a perennial question in all predictive analysis: if we create a forecast model from region A (based on observed outcomes) can we apply the same model in region B, where we do not observe or have limited observations of those outcomes? This question becomes especially pertinent when trying to move from research to operational applications in developing world countries where data are often limited, and new data streams do not occur often (Kebede *et al* 2024, Lee *et al* 2024a).

Most efforts that explore the spatial transferability of models rely on some form of clustering or spatial clustering algorithms (Meyer *et al* 2019, Meyer and Pebesma 2021, Ludwig *et al* 2023). In the context of agriculture, spatial clustering can identify geographically proximate areas with similar climatic, water-use, and soil conditions, crop production patterns, and other factors that influence agricultural production. Cluster-based analyses are a natural starting point for analyzing the spatial transferability of a model. Spatial clustering can be used to identify regions with similar rainfall patterns and soil types, providing some indication of where, and to what extent, a model fit in one region could be applied to another. The advantage of using this approach with EO products is that they tend to have global extents and, thus, there is potential to transfer within a wide spatial extent.

However, EO products are limited in that they cannot capture (though they can be correlated with) key factors that can influence production, including but not limited to different cropping regimes (irrigated vs. non-irrigated), management practices, supply/cost of labor, and other critical socioeconomic factors. In addition, the utility of EO products for forecasting can vary across regions and throughout the growing season (Davenport *et al* 2019, Lee *et al* 2022). For instance, early season precipitation anomalies can indicate opportunities to plant earlier or later, thereby increasing or reducing the productive growth window before the rains cease. Normalized difference vegetation index (NDVI), which roughly estimates photosynthetic activity, may not be very useful during planting and early growth stages, but can be a strong indicator of field productivity during the mid-to-late season. Thus, the spatial transferability of a given model may vary by region, season, and unobserved socioeconomic factors influencing production.

Prior research by Meyer and Pebesma (2021) has proposed examining the area of applicability (**AoA**) of a model based on structurally similar characteristics in the EO predictors and weights based on the model-derived feature importance. They build a metric (known as the Dissimilarity Index) based on the idea that if a variable has a strong contribution to forecast accuracy in one region, then the model should be transferable to other regions where the features of that same variable are similar. We expand on and evaluate this approach in the context of grain yield forecasting in Sub-Saharan Africa (SSA). Specifically, we evaluate an AoA methodology established for generating raster surfaces and apply it to vector-supported grain data. We fit a series of ensemble tree models both within single countries and across multiple sets of countries and then test those models in countries excluded from the training set. We then calculate and decompose AoA metrics and examine several different performance metrics.

We build on the existing literature on spatial transferability and yield forecast models in several ways. We test approaches designed for gridded data, where outcomes are based on continuous latent fields, with vector data, where each data point represents aggregate reports of administrative units in a wide variety of shapes and sizes. We also explore the seasonality of spatial transferability by examining if there are environmental or model characteristics that change the spatial out-of-sample accuracy through the course of the growing season. Finally, we contribute to the broader literature on agricultural yield forecasting by examining what specific factors might make a model fit to one country (or group of countries) more transferable and what environmental characteristics might make a country with little to no agricultural statistics a good candidate for a spatially transferred model. Our objective is to answer the following questions:

1. What are the key characteristics that make a forecast model fit for one set of countries work in another country?
2. Can pooling models across multiple countries provide more accurate out-of-sample estimates than a model fit to one country or district?
3. Does a forecast model fit early in the growing season have the same transferability as a model fit late in the season?

Our paper proceeds as follows: we present a brief overview of the current state of EO-driven yield forecasting in developing countries, as well as efforts to evaluate and maximize the spatial transferability of models. The following section describes the crop yield statistics and EO data we use for our experiments as well as the models, cross-validation structure, and evaluation procedures. Results and discussion are presented in the final two sections.

2. Literature review

There is an increasing number of researchers and institutions using statistical and (or) machine learning methods to forecast crop yields in the developing world (van Klompenburg *et al* 2020). In contrast to deterministic models, these approaches use EOs, agricultural statistics, and other sources of socioeconomic data. For example, in studies of Africa, several authors have combined EO products with statistical methods and/or machine learning to model or predict maize yields or crop cover (Davenport *et al* 2015, 2018, 2019, Lee *et al* 2022, Lobell *et al* 2015, Nakalembe *et al* 2021).

Spatial out-of-sample prediction can generally be divided into two different approaches—kriging (for continuous data) and small area estimation (SAE) (for discrete data). Kriging is a geostatistical interpolation technique that uses the spatial correlation among data points to estimate unknown values at unsampled locations. It assumes that the distance or direction between data points reflects a spatial correlation that can be modeled (typically with a variogram) as an unobserved latent field, allowing for more accurate predictions than traditional linear interpolation (Cressie 1993, Cressie and Wikle 2015). Kriging and related methods are generally used to model latent fields such as precipitation, temperature, elevation, and other geophysical phenomena. When modeling spatially discrete outcomes (such as poverty, employment, and/or demographics) where the underlying spatial process is either unknown or cannot be modeled, the typical approach is to use SAE (Ghosh and Rao 1994, Demombynes *et al* 2007, Tarozzi and Deaton 2007). SAE typically uses some form of linear or non-linear regression on survey data to produce estimates for regions or populations where survey outcomes are unobserved but predictor variables are observed.

Kriging, SAE, and related methods all depend on applying a model based on observed outcomes to a spatial (or spatial-temporal) space where those outcomes are unobserved. The underlying challenge is that models can only be validated within the observed spatial-temporal extent, and it is often unclear how much outside of this observed extent the fitted model will apply.

In Meyer and Pebesma (2021), the authors address the challenge of reliably applying spatial prediction models to areas beyond their training data. They introduce the concept of an 'area of applicability' (AOA), defined as the region where a model's cross-validation error is representative. To delineate the AOA, they propose a 'dissimilarity index' (DI), calculated using the minimum distance to training data in a multidimensional predictor space, with variable weighting based on their importance in the model. This DI helps determine the threshold for the AOA.

The methodology involves standardizing predictor variables, weighting them according to their importance, and then calculating the (weighted) Euclidean distance between data points to compute the DI. The authors tested various DI threshold values against nearly 1000 simulated prediction tasks, ultimately selecting the 0.95 quantile of the DI values from the training data as the threshold. This threshold provides a binary indicator of whether a given region falls inside or outside the AoA while the complementary DI provides a continuous measure (with higher scores being more dissimilar).

3. Data

3.1. Dependent variable

The maize yield data come from official reports issued by the Ministries of Agriculture in Burkina Faso, Kenya, and Malawi and by the FSNAU (Food Security & Nutrition Analysis Unit) in Somalia. The data have been collated by the FEWS NET Data warehouse project that has also performed additional validation and quality control to ensure that the data can be compared across administrative units for the entire period. Table 1 shows the number of administrative units, range of years, and summary yield statistics (across all years and administrative units) for each of the countries, while figure 1 plots average yield for the 10 most recent years.

3.2. Predictor variables

Physical environmental factors in a grain yield forecasting system typically include measures of water availability (precipitation), evaporative demand, crop water-use, and/or photosynthetic activity. Our main predictors are precipitation, evaporative demand, NDVI, cropped area, soil moisture at 5 cm, and soil organic carbon stock at 5 cm. For the EO variables, we use products that measure or model these components, have been used in other yield forecasting and modeling activities, and have a minimum of 20 years of monthly data. We also focus on EO products that are routinely updated on a monthly or sub-monthly basis and thus can be used in an operational forecast setting. The specific predictors are listed in table 2. However, we also use time-invariant products, a crop mask, and estimates of soil water capacity and organic carbon stock. These variables provide contextual information in lieu of spatial dummy variables.

In this paper, we focus on predicting subnational crop yields with 20 or more years of record. The historical dataset is essential to capture infrequent but severe drought events like those linked to El Niño, which tend to occur at intervals of 5–10 years (Timmermann *et al* 1999). We aim to encompass as many of these less-frequent occurrences as possible in both our training and testing datasets.

Figure 2 shows average seasonal patterns in the time-varying variables, stratified by the ranking (low, medium, and high) of yields for each administrative district within their respective countries. Thin lines represent individual districts while thicker lines are aggregate trends. Seasonal progression within the agricultural cycle is delineated into four stages: pre-planting (Pre), early-season (Early), mid-season (Mid), late-season (Late), and post-harvest (Post). We present this figure to highlight similar and divergent seasonal patterns across countries and agricultural regimes (high, medium, and low producers) and because these patterns should be indicative of how transferable a model is (or is not) across countries. We would expect countries and regions with similar seasonal curves for specific variables to have a high potential for spatial transfer. We investigate this more fully with our forecast experiments.

4. Methods

4.1. Model fitting and training

We fit random forest models that predict maize yields in administrative units in four countries (figure 1). Crop yields are observed once per season, but yield forecasting is a concern throughout the season, and updated EO observations are available on a monthly (and sometimes sub-monthly) basis. Thus, we update and re-fit models for each month of the season.

Table 1. Number of administrative units with yield data, first/last year of data, along with mean, standard deviation, and minimum/maximum of reported yields, and total number of observations by country. The final column ('obs. cross-country') refers to the total number of observations from all of the other countries (excluding the one in that row). This is the number of observations used (in the cross-country experiments) to train models that are then tested on that country. Yields are in metric tons per hectare.

Country	# of admin units	First year	Last year	Mean	Std.dev	Min	Max	Obs. (in-country)	Obs. (cross-country)
Burkina Faso	45	1984	2019	1.15	0.51	0.034	2.769	1541	3274
Kenya	47	1982	2019	1.611	0.965	0.004	4.722	1569	3246
Malawi	28	1984	2017	1.398	0.633	0.005	3.378	869	3946
Somalia	38	1995	2021	0.51	0.289	0.02	1.648	836	3979

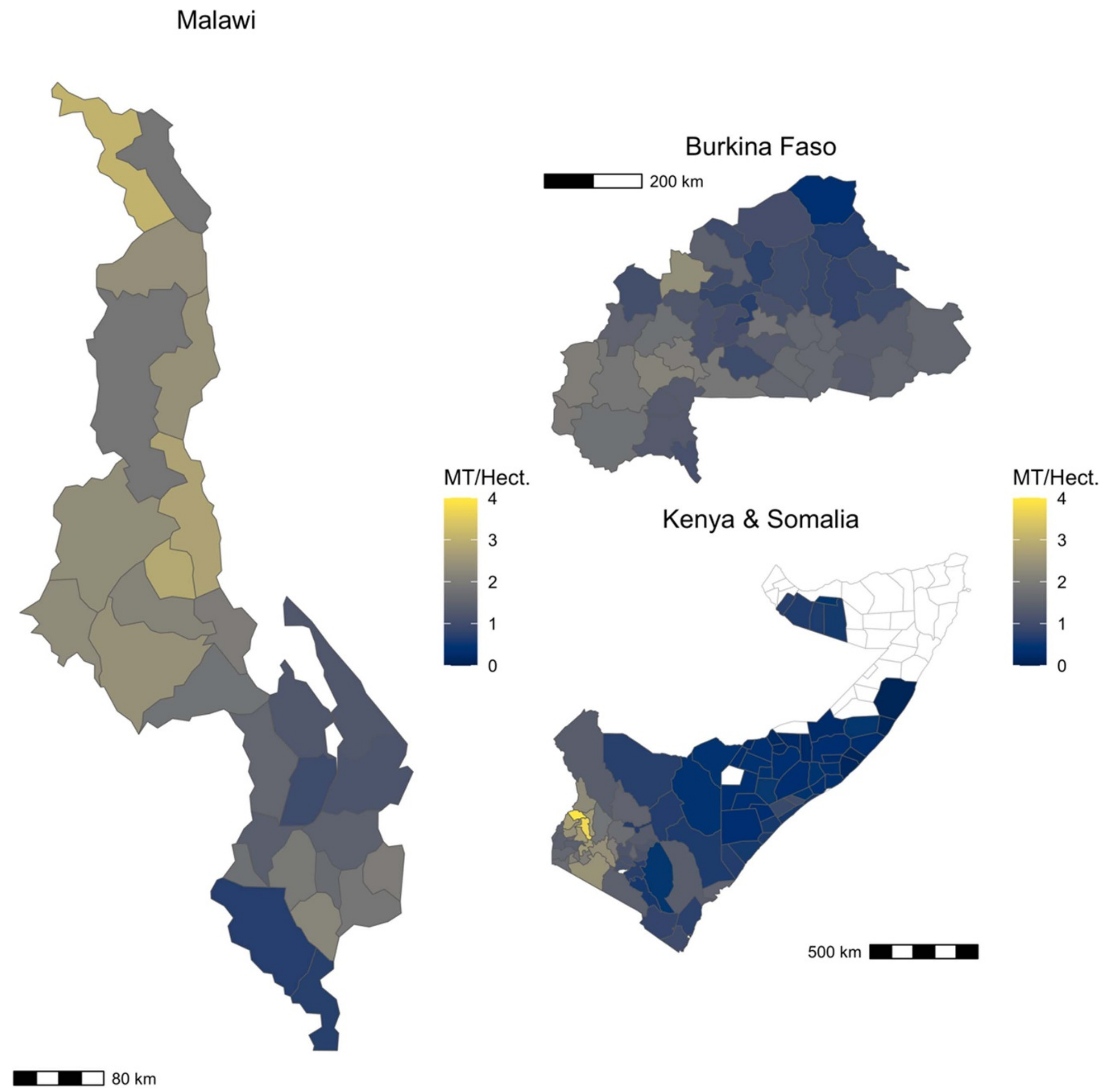

Figure 1. Average maize yields (in metric tons per hectare) for the most recent 10 years of data in each study area. Note that each map is drawn at a different scale (as indicated by separate scale bars).

Prior research suggests that hyper-parameter tuning does not have a substantial influence on the out-of-*spatial-sample predictive* accuracy in random forest models (Schratz *et al* 2019, Meyer and Pebesma 2021, Milà *et al* 2022, Ludwig *et al* 2023). In these cases, out-of-spatial-sample accuracy is more dependent on a model trained using a spatial-cross validation (or in our case, spatial-temporal) scheme that matches the desired prediction conditions (Schratz *et al* 2019, Meyer and Pebesma 2021, Milà *et al* 2022, Ludwig *et al* 2023). We use a spatial-temporal cross-validation scheme (described below) that follows the prediction conditions for an operational yield forecast model. However, to ensure the model generalizes well, avoids over fitting, and is reproducible, we follow the parameter tuning guidelines described in Meyer and Pebesma (2021). Specifically, we use 500 trees in each forest, set *mtry* to be between 2 and the number of predictors (7) and use a grid-search approach to find the optimal parameters. We also follow the guidelines

Table 2. Variables and data sources.

Variable	Product	Spatial resolution temporal extent	Measure
Precipitation	CHIRPS (Funk *et al* 2015) (Climate Hazards Group InfraRed Precipitation with Station data)	0.05° (∼5 km) 1981–Present	Cumulative total since first month of growing season
Evaporative demand	Reference Evapotranspiration (ET_0) monitoring data set (uses MERRA-2 atmospheric reanalysis) (Hobbins *et al* 2016)	0.125° (∼12.5 km) 1981–Present	Cumulative total since first month of growing season
NDVI MAX	AVHRR (Pinzon and Tucker 2014) eVMOD/eVIIRS(Huete *et al* 2002, Jenkerson *et al* 2010)	0.01° (∼1 km) 1981–2002- 375 m 2002–Present[4]	Mean of monthly maximum values since first month of the growing season
Cropland mask	IFPRI-IIASA cropland mask (Fritz *et al* 2015)	0.01° (∼1 km) static data	Proportion of cropland area to total area of polygon (Cropland Percent)
Soil water capacity	Global gridded soil information—SoilGrids (Hengl *et al* 2014)	0.01° (∼1 km), static data, four depth strata including D1 (0–5 cm), D2 (5–15 cm), D3 (15–30 cm), and D4 (30–60 cm).	Soil water capacity at 5 cm
Soil organic carbon stock	Global gridded soil information—SoilGrids (Hengl *et al* 2014)	0.01° (∼1 km) static data, four depth strata including D1 (0–5 cm), D2 (5–15 cm), D3 (15–30 cm), and D4 (30–60 cm).	Soil organic carbon stock (Soil OCS) at 5 cm

from the bootstrapping literature and sample with replacement with the sample size set to the size of the training sample (Chernick 2007, Cameron *et al* 2008, Burridge and Fingleton 2010).

We use two different training and validation approaches.

1. Training models on one country and testing on others. In these experiments we fit a model on one country and test the transferability to other countries. We use spatial-temporal training folds, leaving out years and blocks of administrative units of the country we train on. We then test the transferability of the models on all countries not included in the training data.
2. Training models on all but one country and testing on the holdout country. Again, we use spatial-temporal training folds, but we hold out an entire country and year during each iteration. We then test the transferability of the model on the country not included in the training data.

By systematically holding out spatial and temporal blocks, we ensure that the cross-validation process accurately reflects the model's performance in out-of-sample spatial forecasts, approximating the model's transferability across different spatial scales and environmental contexts. We evaluate the transferability of models by predicting yields for all months and years in the holdout countries. All accuracy measures we present below are for out-of-sample observations (excluded from the training data). We calculate the mean absolute percent error to evaluate model skill and examine how model skill correlates with seasonal and environmental factors as well as the dissimilarity index introduced by Meyer and Pebesma (2021).

5. Results

5.1. Variable importance

Variable importance measures are shown in figure 3, stratified by variable (panels), the country (or countries) the model was trained on, and the period in the season when the model was run (*x*-axis). We include these results to both highlight heterogeneity across

[4] We blended two distinct normalized difference vegetation index (NDVI) datasets to achieve a continuous time series analysis spanning from 1981 to 2021. The NOAA AVHRR NDVI (1982–2002) and USGS EROS eVMOD/eVIIRS (2002–2021) datasets were integrated, applying bias correction to align means and standard deviations. The eVMOD NDVI dataset, specifically, represents a refined version of the USGS EROS MODIS (eMODIS) NDVI, which has been adjusted to align with the eVIIRS NDVI through geometric mean regression.

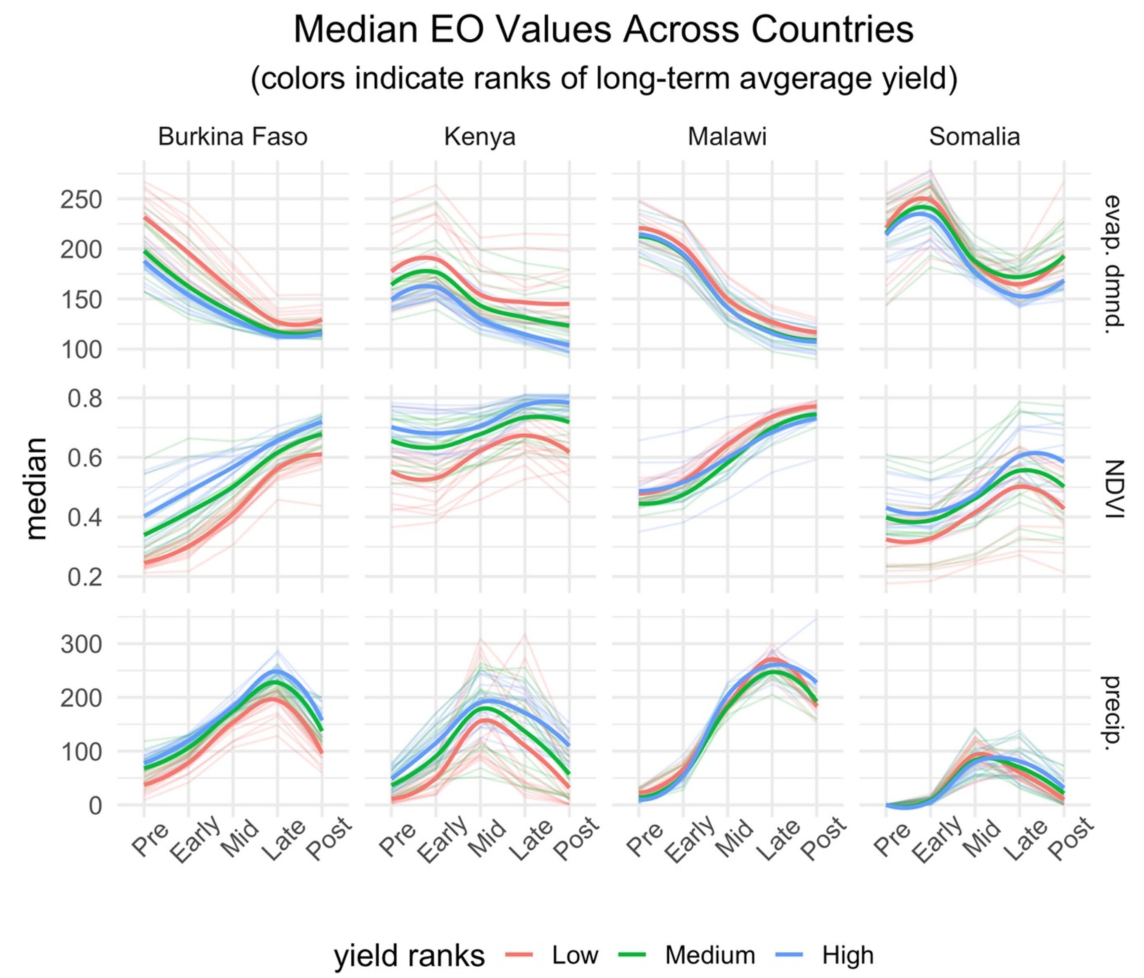

Figure 2. Median Earth Observations (y-axis) during stages of the growing season (x-axis) and countries (columns). Thick lines are country level trends and thin lines are individual districts. Colors indicate long-term average yield ranks relative to other districts in the same country. Precipitation and evaporative demand are measured in millimeters while NDVI is an index scaled between −1 and 1.

optimal model fits for each country and because similarity in features and the importance of those features are a critical component of the DI index developed by Meyer *et al* (2019). If the index is an accurate measure of transferability, then countries with similar patterns in the features (figure 2) and variable importance measures should be indicative of spatially transferable models.

In general, we find that variable importance fluctuates across different countries and time frames, suggesting that the influences of environmental and agricultural factors on model performance are different in each region. Evaporative demand shows moderate importance in the early season, with distinct peaks in Kenya during the middle of the season. NDVI displays high relative importance for Malawi, while precipitation was important throughout the season for Kenya and in later stages for Burkina Faso and Somalia. The time-invariant variables (percent crop land, soil water capacity, and soil organic carbon stock) had the highest relative importance because they vary neither through time nor through the season and likely capture numerous geographic and spatial effects. Percent crop area maintains a fairly consistent pattern across the countries, with an observable increase in importance in the late time frame for Somalia. We also emphasize here that variable importance measures do not necessarily reflect actual physical mechanisms at play but simply provide diagnostics of how the model is functioning and training the features available.

5.2. Distribution of MAPE scores across seasonal stages and long-term average rainfall

Our first set of forecast experiment results focus on forecast accuracy across different stages of the growing season. Figure 4 shows the distribution of mean absolute percentage error (MAPE) scores for models trained on data from different countries, distinguished by the average seasonal rainfall of the area tested. We stratify the data in this way because we expect variation in seasonal forecast skill (and the

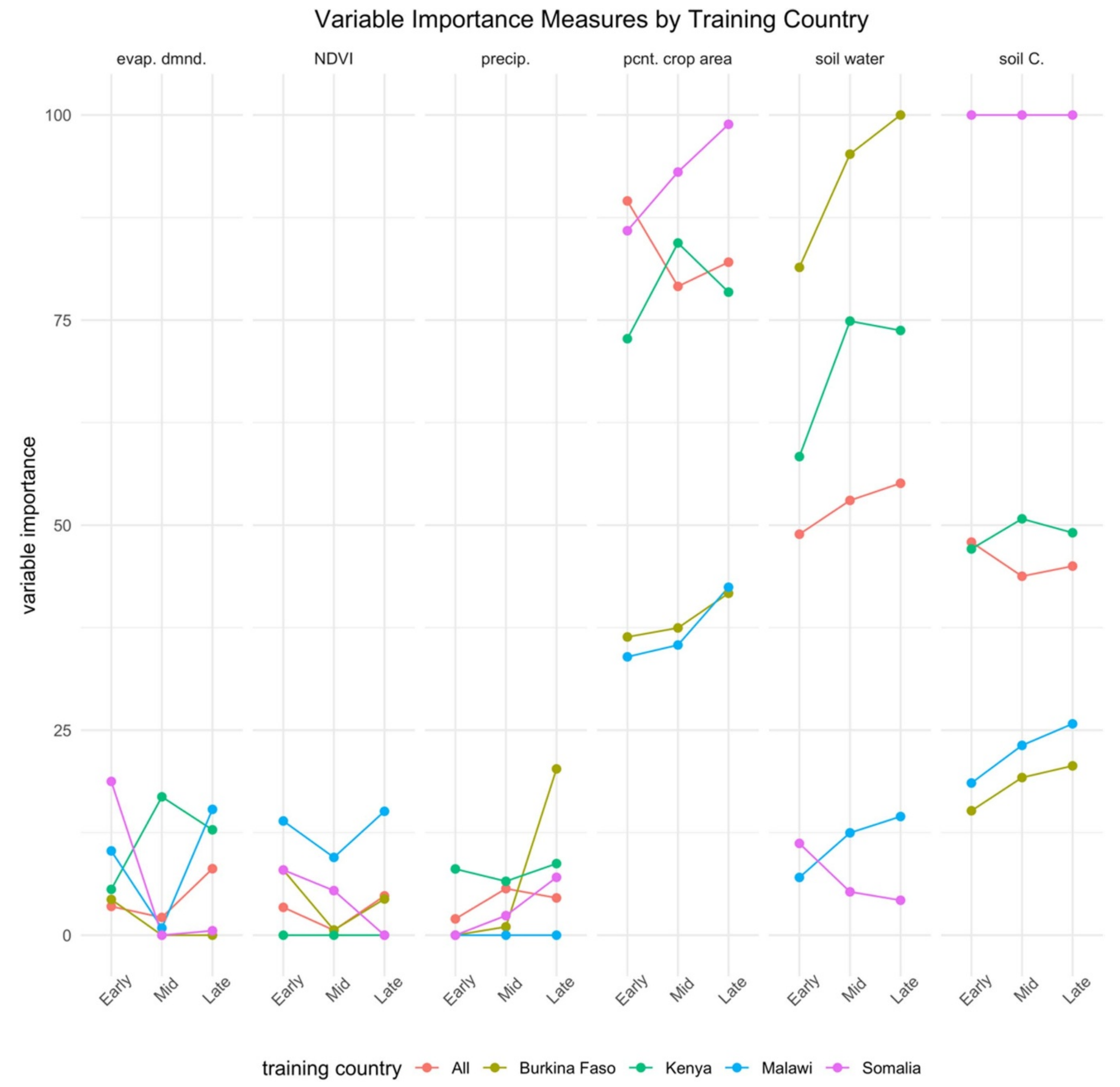

Figure 3. Variable importance measures (y-axis) from models trained on sets of countries (colors) during different periods in the season (x-axis). Note that the values for 'All' are averaged across all models where one country was held out.

transferability of models) across drier (arid to semi-arid with <500 mm annual precipitation) versus wetter regions (500–750 mm and >750 mm). Generally, we notice distinct patterns in accuracy across drier (less accurate) or wetter (more accurate) regions, but we do not see strong differentiation in patterns across the growing season. There are some exceptions to this. During the early phase, the model trained on Kenya exhibits a peak within the 500–750 mm range (green), indicating a clustering of lower MAPE scores. Conversely, the model trained on Malawi displays a blue peak in the late phase, indicating higher forecast accuracy for regions receiving more than 750 mm of rainfall. The model trained on Somalia reveals broad distributions across all phases and precipitation ranges. Models trained on Burkina Faso data tend to show the least variation in MAPE scores across different seasonal stages. Overall, the distributions convey that the predictability of the models varies not only by training set but also aligns with the average precipitation of the test region.

5.3. Relationship between forecast skill and average annual precipitation

We now examine the relationship between forecast skill and average total precipitation across varying within-country ranks of average yields (low, medium, high, relative to other regions in the same country). We stratify the data in this way because we expect forecast skill (and the transferability of models) to vary across low and high yield areas, as these areas might represent both different climate and agricultural (rain fed or irrigated) regimes. We also expect models trained on drier (arid to semi-arid with <500 mm annual precipitation), lower yield areas to perform better when tested on similar regions and vice versa. In figure 5, we plot MAPE scores (y-axis) versus average annual precipitation

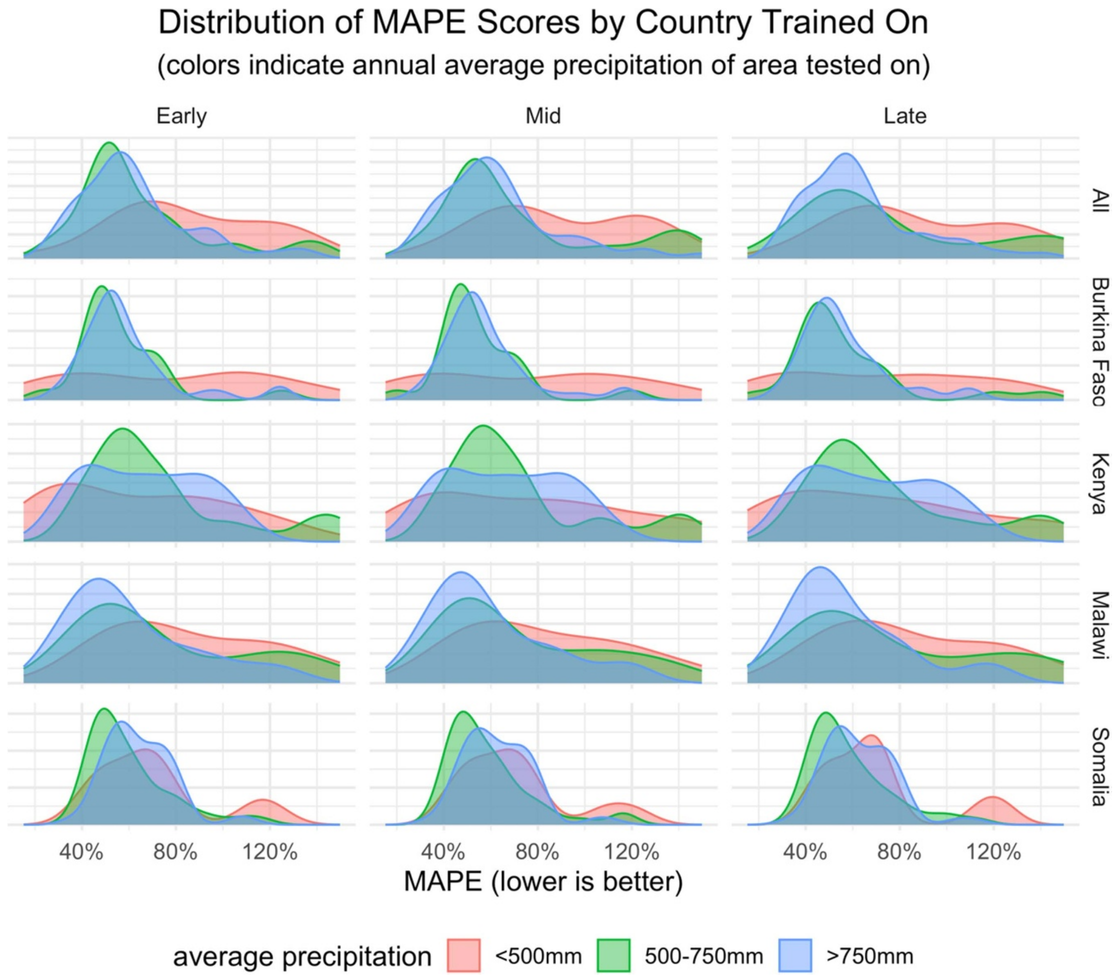

Figure 4. Distribution of mean absolute percentage error (MAPE) scores for models trained on data from different countries, distinguished by the average annual rainfall of the area tested. The density distributions for MAPE scores are segmented into three phases: early, mid, and late. Each panel represents one of the countries—Burkina Faso, Kenya, Malawi, Somalia—or an aggregate of all ('All'), with distributions colored based on three average precipitation ranges: red for less than 500 mm, green for 500–750 mm, and blue for more than 750 mm.

where a model was tested (x-axis). Data points are color-coded to represent the countries the models were trained on, including an aggregated dataset ('All'), Burkina Faso, Kenya, Malawi, and Somalia, while shapes signify the country where a model is tested.

When observing the low yield rank, there is an evident increase in MAPE as the average total rainfall rises, peaking around 800 mm, then sharply declining; however, the overall relationship is quite noisy. The medium yield rank areas also display a dip in MAPE around the 800 mm mark, indicating a positive relationship between forecast accuracy and wetter regions at this transition point. Higher yielding areas, in contrast, show a flatter error rate across varying precipitation levels, with only slight fluctuations. Collectively, these results suggest that there is some relationship between structural similarities in the predictors (in this case precipitation) and predictive accuracy- we further investigate this relationship in figure 6.

5.4. Relationship between dissimilarity index and forecast accuracy

Figure 6 further disaggregates the results shown in figure 5 by showing separate panels based on the training and test areas. We focus only on districts where the MAPE score is ⩽45% and show the DI scores (from Meyer and Pebesema 2021) for the test areas. To preserve the legibility of the figure, we only show plots that have the expected negative correlation between MAPE scores and average precipitation (supplemental figure S1 shows this relationship for all areas). The regression lines shown in figure 6 indicate a decrease in MAPE with an increase in average total precipitation, which aligns with the hypothesis that models are more accurate in conditions that resemble

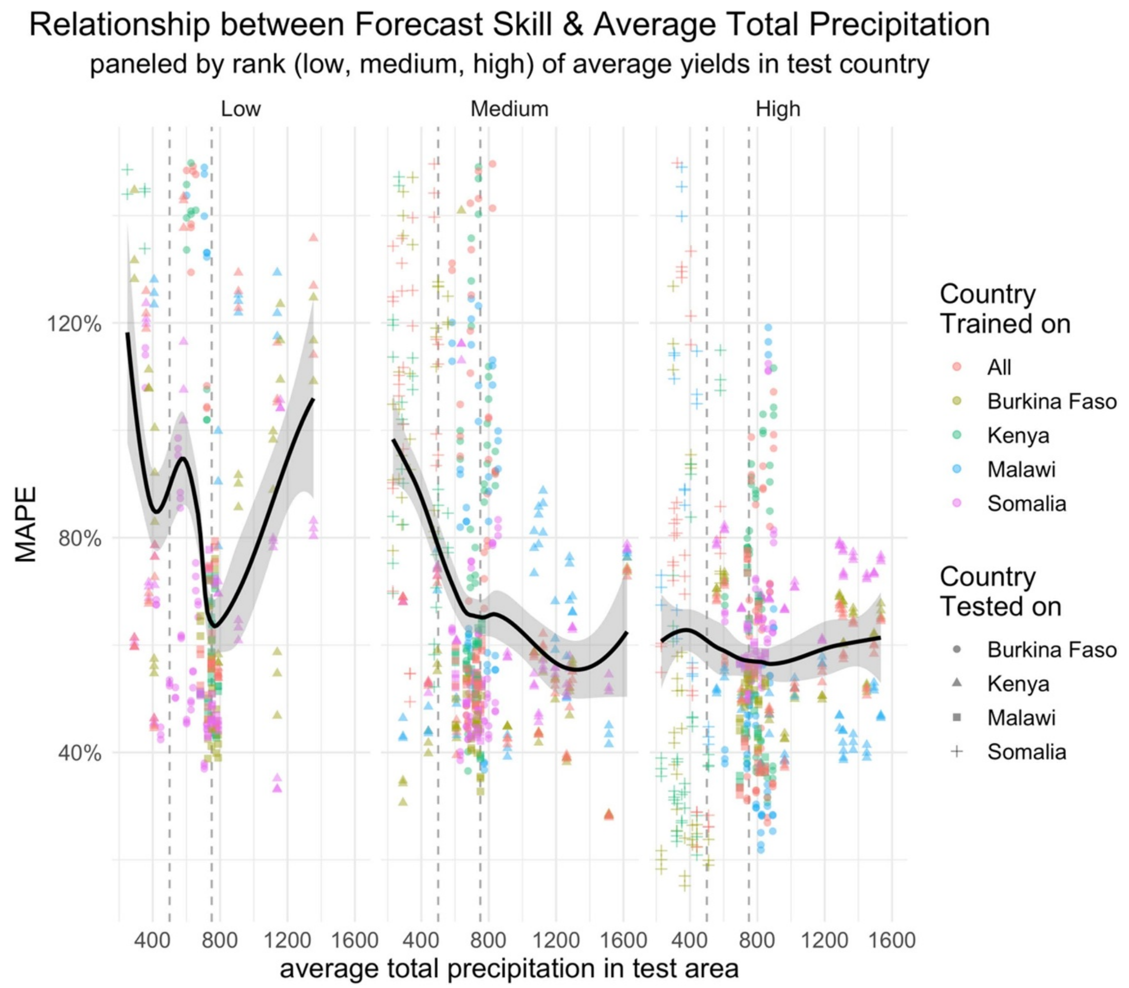

Figure 5. Relationship between forecast skill (y-axis, lower is better), average annual precipitation in an administrative unit (x-axis), and within country yield ranks (panels). Data points are color-coded to represent the countries the models were trained on, including an aggregated dataset ('All'), Burkina Faso, Kenya, Malawi, and Somalia. Shapes represent the country where a model is tested. Vertical dashed lines show thresholds for Arid-Semi-Arid regions (<500 mm) and for where rainfed maize production is ideal (>750 mm).

their training environment, as indicated by lower DI scores.

6. Discussion and conclusion

We examine the spatial transferability of EO-driven yield forecast models within SSA. Given the expansive reach of EO data contrasted with the more limited scope of high-quality agricultural survey data, we test whether a forecast model developed in one region can be validly applied to another region with scarce or no data.

Forecast accuracy, measured by MAPE scores, aligns with the expected trend of higher accuracy in regions with precipitation patterns similar to the training data. These patterns are not uniform across high, medium, and low yielding areas, pointing to the nuanced nature of spatial transferability. In contrast to the yield ranks, we do not find widely varying accuracy across the season, indicating that a model trained and transferred on early season data will likely be equally good or bad when used on later season data.

In examining the relationship between forecast skill and average total precipitation across yield ranks, we note that although the relationship can be noisy, certain trained/tested combinations (trained on All and tested on Burkina Faso and Kenya; trained on Burkina Faso and tested on Kenya, Malawi, and Somalia; trained on Kenya and tested on Burkina Faso; and trained on Malawi and tested on Burkina Faso and Kenya) exhibit the hypothesized negative correlation between MAPE and precipitation, particularly when models are applied to regions with similar climates to those they were trained on. Time-invariant variables like percent crop land, soil water capacity, and soil organic carbon stock, which should capture numerous geographic and spatial effects, also likely play a crucial role in model transferability.

The results indicate that pooling models across multiple countries can indeed provide more accurate out-of-sample estimates compared to models fitted

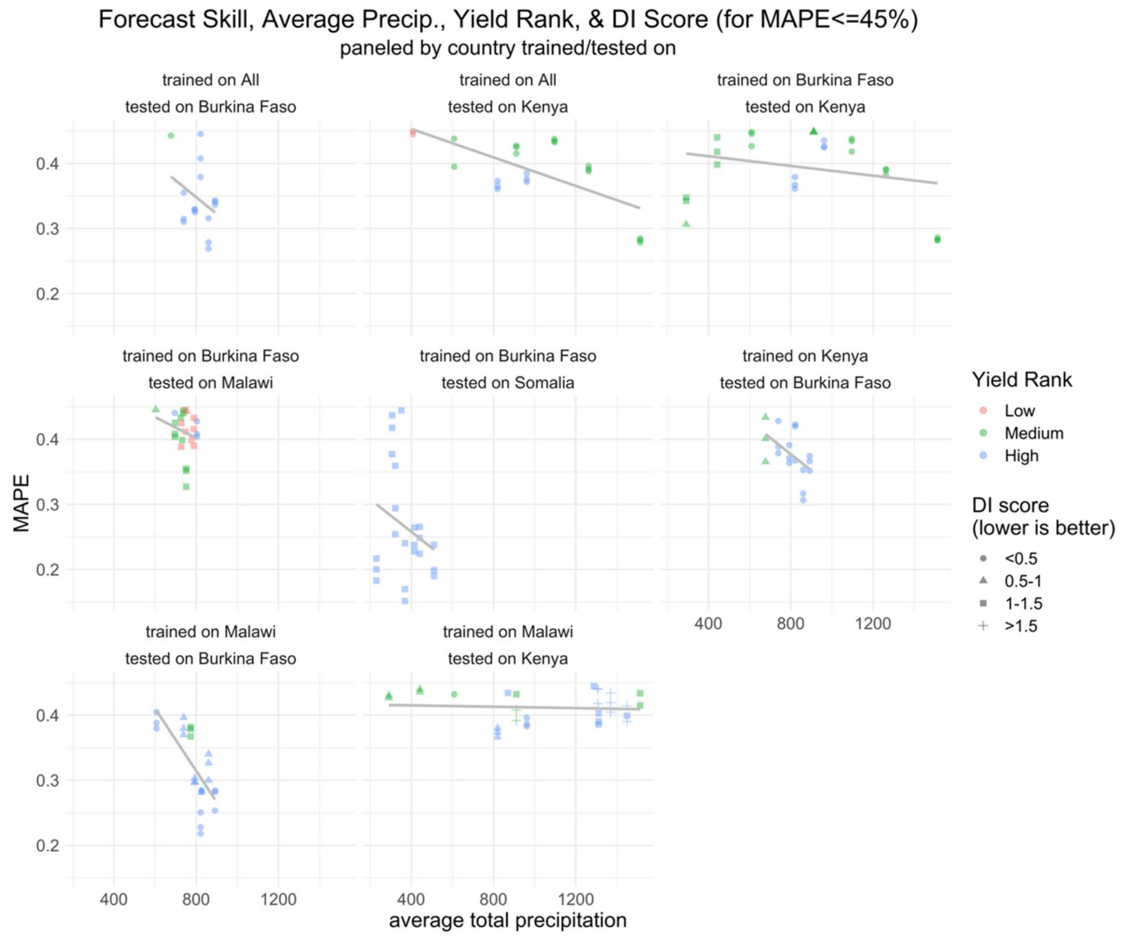

Figure 6. Disaggregation of the interplay among forecast skill, average total precipitation, yield rank, and the dissimilarity index (DI), filtered to include only results where the mean absolute percentage error (MAPE) is less than or equal to 45%. The yield ranks are the same as shown in figure 5 and are relative to other regions within the country where the model is tested.

to a single country or district. This improved accuracy is attributed to the broader range of environmental and agricultural data that the models can learn from, allowing them to capture a wider array of conditions that might be present in the test regions. The pooled data sets provide a more robust and generalized model that can better accommodate the variability encountered in different geographical areas.

Data availability statement

The data cannot be made publicly available upon publication because they are owned by a third party and the terms of use prevent public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgment

This work is supported by USGS/USAID cooperative Grants #G22AC00565-00, #72DFFP19CA00001, and #444032-22603

Earlier versions of this Paper were presented at the American Geophysical Union meetings (AGU 2022), the European Geophysical Union meetings (EGU 2023), and the 2024 Western Regional Science Association meetings (WRSA 2024). We thank our WRSA discussant, Shiyan Zhang, for her comments on this paper. We also thank Harikshan Jayanthi and Kul Khand for their reviews of this paper. Finally we thank two anonymous reviewers from ERL. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

ORCID iDs

F Davenport https://orcid.org/0000-0003-2409-768X
D Lee https://orcid.org/0000-0001-5438-903X
S Shukla https://orcid.org/0000-0003-0077-6733

References

Becker-Reshef I, Vermote E, Lindeman M and Justice C 2010 A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data *Remote Sens. Environ.* **114** 1312–23

Burridge P and Fingleton B 2010 Bootstrap inference in spatial econometrics: the J-test *Spatial Econ. Anal.* **5** 93–119

Cameron A C, Gelbach J B and Miller D L 2008 Bootstrap-based improvements for inference with clustered errors *Rev. Econ. Stat.* **90** 414–27

Chernick M R 2007 *Bootstrap Methods: A Guide for Practitioners and Researchers* (Wiley-Interscience)

Cressie N 1993 *Statistics for Spatial Data* (Wiley)

Cressie N and Wikle C K 2015 *Statistics for Spatio-temporal Data* (Wiley)

Davenport F M, Harrison L, Shukla S, Husak G, Funk C and McNally A 2019 Using out-of-sample yield forecast experiments to evaluate which earth observation products best indicate end of season maize yields *Environ. Res. Lett.* **14** 124095

Davenport F M, Shukla S, Turner W, Funk C, Krell N, Harrison L, Husak G, Lee D and Peterson S 2021 Sending out an SOS: using start of rainy season indicators for market price forecasting to support famine early warning *Environ. Res. Lett.* **16** 1748–9326

Davenport F, Funk C and Galu G 2018 How will East African maize yields respond to climate change and can agricultural development mitigate this response? *Clim. Change* **147** 491–506

Davenport F, Husak G and Jayanthi H 2015 Simulating regional grain yield distributions to support agricultural drought risk assessment *Appl. Geogr.* **63** 136–45

Demombynes G, Elbers C, Lanjouw J O and Lanjouw P 2007 How good a map? Putting small area estimation to the test *Policy Research Working Paper Series* (The World Bank)

FAO 2016 Crop yield forecasting: methodological and institutional aspects current practices from selected countries (Belgium, China, Morocco, South Africa, USA) with a focus on AMIS crops (maize, rice, soybeans and wheat) (FAO)

Fritz S *et al* 2015 Mapping global cropland and field size *Glob. Change Biol.* **21** 1980–92

Funk C, Peterson P, Landsfeld M, Pedreros D, Verdin J, Shukla S, Husak G, Rowland J, Harrison L and Hoell A 2015 The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes *Sci. Data* **2** 1–21

Ghosh M and Rao J N K 1994 Small area estimation: an appraisal *Stat. Sci.* **9** 55–76

Hengl T *et al* 2014 SoilGrids1km—global soil information based on automated mapping *PLoS One* **9** e105992

Hobbins M T, Wood A, McEvoy D J, Huntington J L, Morton C, Anderson M and Hain C 2016 The evaporative demand drought index. Part I: linking drought evolution to variations in evaporative demand *J. Hydrometeorol.* **17** 1745–61

Huete A, Didan K, Miura T, Rodriguez E P, Gao X and Ferreira L G 2002 Overview of the radiometric and biophysical performance of the MODIS vegetation indices *Remote Sens. Environ.* **83** 195–213

Jenkerson C B, Maiersperger T and Schmidt G 2010 eMODIS: a user-friendly data source *Report* 2010–1055*; Open-File Report* (USGS Publications Warehouse) 10.3133/ofr20101055

Johnson D M 2014 An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yields in the United States *Remote Sens. Environ.* **141** 116–28

Kebede E A *et al* 2024 Assessing and addressing the global state of food production data scarcity *Nat. Rev. Earth Environ.* **5** 295–311

Kouadio L, Newlands N K, Davidson A, Zhang Y and Chipanshi A 2014 Assessing the performance of MODIS NDVI and EVI for seasonal crop yield forecasting at the ecodistrict scale *Remote Sens.* **6** 10193–214

Krell N, Davenport F, Peterson S, Shukla S, Husak G J, Turner W, Funk C C and Caylor K K 2019 To what extent does climate variability explain farmers' planting decisions in central Kenya *AGU Fall Meeting Abstracts*

Lee D, Anderson W, Chen X, Davenport F, Shukla S, Sahajpal R, Budde M, Rowland J, Verdin J and You L 2024a HarvestStat Africa–harmonized subnational crop statistics for sub-Saharan Africa

Lee D, Davenport F, Shukla S, Husak G, Funk C, Budde M, Rowland J and Verdin J 2024b Contrasting performance of panel and time-series models for subnational crop forecasting in sub-Saharan Africa *Agric. For. Meteorol.* accepted (https://doi.org/10.1016/j.agrformet.2024.110213)

Lee D, Davenport F, Shukla S, Husak G, Funk C, Harrison L, McNally A, Rowland J, Budde M and Verdin J 2022 Maize yield forecasts for sub-Saharan Africa using Earth observation data and machine learning *Glob. Food Secur.* **33** 100643

Lobell D B, Thau D, Seifert C, Engle E and Little B 2015 A scalable satellite-based crop yield mapper *Remote Sens. Environ.* **164** 324–33

Ludwig M, Moreno-Martinez A, Hölzel N, Pebesma E and Meyer H 2023 Assessing and improving the transferability of current global spatial prediction models *Glob. Ecol. Biogeogr.* **32** 356–68

Meyer H and Pebesma E 2021 Predicting into unknown space? Estimating the area of applicability of spatial prediction models *Methods Ecol. Evol.* **12** 1620–33

Meyer H, Reudenbach C, Wöllauer S and Nauss T 2019 Importance of spatial predictor variable selection in machine learning applications—moving from data reproduction to spatial prediction *Ecol. Modelling* **411** 108815

Milà C, Mateu J, Pebesma E and Meyer H 2022 Nearest neighbour distance matching leave-one-out cross-validation for map validation *Methods Ecol. Evol.* **13** 1304–16

Nakalembe C *et al* 2021 A review of satellite-based global agricultural monitoring systems available for Africa *Glob. Food Secur.* **29** 100543

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

Pinzon J E and Tucker C J 2014 A non-stationary 1981–2012 AVHRR NDVI3g time series *Remote Sens.* **6** 6929–60

Schauberger B, Jägermeyr J and Gornott C 2020 A systematic review of local to regional yield forecasting approaches and frequently used data resources *Eur. J. Agron.* **120** 126153

Schratz P, Muenchow J, Iturritxa E, Richter J and Brenning A 2019 Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data *Ecol. Modelling* **406** 109–20

Shukla S, Husak G, Turner W, Davenport F, Funk C, Harrison L and Krell N 2021 A slow rainy season onset is a reliable harbinger of drought in most food insecure regions in sub-Saharan Africa *PLoS One* **16** e0242883

Tarozzi A and Deaton A 2007 Using census and survey data to estimate poverty and inequality for small areas *Working Papers* (Princeton University, Department of Economics, Industrial Relations Section)

Timmermann A, Oberhuber J, Bacher A, Esch M, Latif M and Roeckner E 1999 Increased El Niño frequency in a climate model forced by future greenhouse warming *Nature* **398** 694–7

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

Zhang Y, Chipanshi A, Daneshfar B, Koiter L, Champagne C, Davidson A, Reichert G and Bédard F 2019 Effect of using crop specific masks on earth observation based crop yield forecasting across Canada *Remote Sens. Appl.* **13** 121–37