Skillful U.S. Soy Yield Forecasts at Presowing Lead Times

SEM VIJVERBERG, RAED HAMED, AND DIM COUMOU

Department of Water and Climate Risk, Institute for Environmental Studies (IVM), Vrije Universiteit Amsterdam, Amsterdam, Netherlands
Royal Netherlands Meteorological Institute (KNMI), De Bilt, Netherlands

ABSTRACT: Soy harvest failure events can severely impact farmers, insurance companies, and raise global prices. Reliable seasonal forecasts of misharvests would allow stakeholders to prepare and take appropriate early action. However, especially for farmers, the reliability and lead time of current prediction systems provide insufficient information to justify within-season adaptation measures. Recent innovations increased our ability to generate reliable statistical seasonal forecasts. Here, we combine these innovations to predict the 1–3 poor soy harvest years in the eastern United States. We first use a clustering algorithm to spatially aggregate crop producing regions within the eastern United States that are particularly sensitive to hot-dry weather conditions. Next, we use observational climate variables [sea surface temperature (SST) and soil moisture] to extract precursor time series at multiple lags. This allows the machine learning model to learn the low-frequency evolution, which carries important information for predictability. A selection based on causal inference allows for physically interpretable precursors. We show that the robust selected predictors are associated with the evolution of the horseshoe Pacific SST pattern, in line with previous research. We use the state of the horseshoe Pacific to identify years with enhanced predictability. We achieve high forecast skill of poor harvests events, even 3 months prior to sowing, using a strict one-step-ahead train-test splitting. Over the last 25 years, when the horseshoe Pacific SST pattern was anomalously strong, 67% of the poor harvests predicted in February were correct. When operational, this forecast would enable farmers to make informed decisions on adaptation measures, for example, selecting more drought-resistant cultivars or change planting management.

SIGNIFICANCE STATEMENT: If soy farmers would know that the upcoming growing season will be hot and dry, they could decide to take anticipatory action to reduce losses, that is, buy more drought resistant soy cultivars or change planting management. To make such decisions, farmers would need information even prior to sowing. On these very long lead times, a predictable signal can emerge from low-frequency processes of the climate system that can affect surface weather via teleconnections. However, traditional forecast systems are unable to make reliable predictions at these lead times. In this work, we used machine learning techniques to train a forecast model based on these low-frequency components. This allowed us to make reliable predictions of poor harvest years even 3 months prior to sowing.

KEYWORDS: Atmosphere-ocean interaction; Rossby waves; Pacific decadal oscillation; Sea surface temperature; Seasonal forecasting; Machine learning

1. Introduction

Seasonal forecasts of U.S. soy yield play a crucial role in the decision making of numerous stakeholders (Klemm and McPherson 2017). Reliable yield forecasts can improve crop management of local farmers and inform (non)governmental organizations on total supply and expected prices (Basso and Liu 2019). Monetary value of soy is ranked first among all staple crops, and the United States supplies one-third of globally traded soybean (Jin et al. 2017), making forecasts highly relevant for commodity traders (Torreggiani et al. 2018). The ongoing increase in soy demand is expected to continue in the future (Fehlenberg et al. 2017), while on the other hand, climate change is expected to threaten U.S. production by increasing average and extreme temperatures (Dirmeyer et al. 2013; Winter et al. 2015). Reliable and timely predictions can mitigate the impacts from climate extremes (WMO 2020; Alley et al. 2019) and are expected to be the most cost-efficient way to increase resilience against the projected impacts of climate extremes (Mbow et al. 2019). The most frequently used adaptation measure by farmers is crop or cultivar selection and adjusting planting timing and/or management (Crane et al. 2010). With reliable and timely forecasts, farmers could 1) better manage irrigation schedules (Villani et al. 2021), 2) buy insurance against crop failure (Li et al. 2019), 3) lower the sowing density (Carter et al. 2018; Lobell et al. 2020, 2014), 4) decide to only plant in lower (i.e., wetter) altitude areas (Crane et al. 2010), or 5) decide to order more drought resistant crops or soy cultivars (normally done already in January/February) (Dong et al. 2019; Arya et al. 2021; Crane et al. 2010). Commodity traders can buy the soy seasons prior to the harvesting period in October at an expected future price, which shifts the risk of future price fluctuations from the farmer to the trader (Bhardwaj et al. 2015). Thus, also traders have a key interest in information on expected yields already several seasons...
before harvest time. Knowing the risk well in advance enables both farmers and traders to make more informed decisions. However, to the best of our knowledge, skillful predictions for harvest failures at lead times before planting in May currently do not exist (Basso and Liu 2019).

Current operational forecasts are based upon surveys, which rely on local observations by experts (Schnepf 2017; National Agricultural Statistics Service 2012). Although these survey-based forecasts can be skillful (Beguería and Maneta 2020), a logical consequence is that they are done during the growing season, seriously limiting adaptation options. Moreover, such surveys do not take into account long-range weather forecasts or any other relevant climatic information (National Agricultural Statistics Service 2012).

Recent studies have shown that U.S. soybean production is vulnerable to the combined effect of hot and dry weather conditions in July and August (Ortiz-Bobea et al. 2019; Haqiqi et al. 2021; Gouart et al. 2021), particularly in the mid- to southern producing regions (Hamed et al. 2021). Extreme conditions (defined by the 95th percentile of temperature and 5th percentile of soil moisture) reduce soy yields by about two standard deviations. This crop sensitivity to hot–dry conditions is 4 times larger compared to hot conditions alone and 3 times larger compared to dry conditions alone (Hamed et al. 2021). The trend in hot–dry conditions seem to have increased over time and is expected to further increase with future climate change (Hamed et al. 2021; Zscheischler and Seneviratne 2017).

To inform stakeholders on the hot–dry hazards, dynamical seasonal forecast models can be used in isolation, or they can be combined with a crop simulation model to predict end-of-season yields. However, the dynamical forecast models do not have sufficient long-lead skill in the United States that is necessary to give warning well in advance (i.e., 3–4 months ahead of sowing) (Kirtman et al. 2014; Jong et al. 2021). On these seasonal time scales, a predictable signal can emerge from low-frequency processes of the climate system that can affect surface weather via teleconnections (Krishnamurthy 2019). However, this predictable signal is generally underestimated in dynamical models (Di Capua et al. 2021; Vijverberg et al. 2020; Merryfield et al. 2020; National Academies of Sciences, Engineering, and Medicine 2016; Scaife and Smith 2018). The poor skill of dynamical seasonal forecasts (Ramírez-Rodrigues et al. 2016) combined with imperfect crop simulation models generates a cascade of uncertainty making the approach unsuited for presowing harvest predictions (Brown et al. 2018; Iizumi et al. 2018).

Machine learning techniques have the potential to circumvent the problem of low signal-to-noise ratios of dynamical models by directly learning from observations. Of course, this is based on the premise that there is inherent predictability in the system. A first essential step is to minimize the unpredictable noise in the target time series, which generally involves spatial and temporal averaging (Krishnamurthy 2019). Second, dimensionality reduction methods are often needed to extract the signal(s) from relevant precursor datasets. The traditional approach is to use the known climate indices, since those capture the first-order low-frequency processes in the climate system [e.g., the Niño-3.4 index to describe El Niño–Southern Oscillation (ENSO)]. However, these climate indices can easily miss important, more detailed, information relevant for a specific target variable of interest, as they can oversimplify the state of a complex dynamical system to a single number (Vijverberg and Coumou 2022). For example, the North Pacific sea surface temperature (SST) is known to affect surface weather in the eastern United States (Liu et al. 2006; McKinnon et al. 2016), and correlations are indeed found with the Pacific decadal oscillation (PDO) index (Kurtzman and Scanlon 2007; Yu and Zwiers 2007). However, the PDO pattern is designed to maximize the explained variability over the entire North Pacific, while ostensibly only a subregion of the Pacific is physically connected to U.S. surface weather via the forcing of atmospheric Rossby waves (Vijverberg and Coumou 2022).

Recent studies have shown that eastern U.S. July–August temperature is well predictable by a horseshoe-like SST pattern in the North Pacific when using machine learning techniques (McKinnon et al. 2016; Vijverberg et al. 2020; Vijverberg and Coumou 2022). Eastern U.S. heatwaves are predictable up to 50 days lead time (Vijverberg et al. 2020), and the July–August mean temperature is highly predictable by the winter-to-spring horseshoe SST pattern (Vijverberg and Coumou 2022). The horseshoe North Pacific SST pattern resembles the Pacific decadal oscillation pattern (Newman et al. 2016) and therefore shows approximately similar decadal variability (Vijverberg and Coumou 2022). Still, there is a lot of variability occurring at inter- and intra-annual time scales. On seasonal time scales, the horseshoe-like SST pattern can force an arcing Rossby wave over the North American continent in summer, which subsequently leads to more persistent and/or frequent high pressure systems over the mid- and eastern United States. High pressure systems are associated with reduced rainfall and higher surface temperature, thereby increasing the risk of hot–dry weather. Besides the ocean–atmosphere forcing in summer, the winter–spring atmosphere–ocean forcing, and two-way ocean–atmosphere feedbacks, also play a role in strengthening the horseshoe-like SST pattern (Vijverberg and Coumou 2022). This strengthening is important for predictability, since during both persistent and anomalous midlatitude SST states there is a stronger atmospheric response (Ferreira and Frankignoul 2005). Hence, during persistent and anomalous horseshoe Pacific states, a window of predictability exists (Mariotti et al. 2020), since this higher signal results in a pronounced increase in forecast skill for eastern U.S. temperatures (Vijverberg and Coumou 2022).

Feeding a statistical model with information from multiple lags (instead of only the most recent lag) can often help to improve forecast skill (Vijverberg and Coumou 2022; Switanek et al. 2020). For example, considering the past evolution informs on the life cycle of ENSO (if El Niño is in a growing, decaying, or persistent phase) and thus contains more information compared to only a single snapshot, which is important for forecast skill.

Here, we aim at forecasting poor harvest years of U.S. soybean by combining three recent insights: 1) the mid-to-south producing regions are vulnerable to hot and dry conditions in summer (Hamed et al. 2021), and 2) seasonal July–August
temperatures are well predictable at seasonal time scales in years with a pronounced horseshoe Pacific SST state (Vijverberg and Coumou 2022), and 3) using features at multiple lags can further boost forecast skill (Switanek et al. 2020; Vijverberg and Coumou 2022). We investigate the potential of a forecasting system that can inform stakeholders on the risk of a poor harvest. Hence, we aim at directly predicting impact, defined as poor soy yield years over an aggregated domain (see method section 2b). Building upon state-of-the-art data-driven techniques, we introduce a new framework that applies a causal inference-based method to select specific precursors to reduce overfitting and improve interpretability and reliability (Fig. 1). Combining these new insights and techniques allows us to achieve high forecast skill for yield in the mid-to-southern producing region already in February, 8 months prior to the harvest period, and 3 months before sowing.

2. Methods

a. Data

For the precursor datasets, we use ERA5 monthly mean sea surface temperature (10°S–60°N) and the volumetric soil water layer (m³ m⁻³) (0–7-cm depth, 135°–60°W, 10°S–60°N) spanning from 1950 up to 2019, both on a 1.0° spatial resolution. The seasonal cycle, calculated as the multiyear mean per month, is removed from the data and subsequently we remove the linear trend. Finally, the SST data are aggregated to 2-month means. Since we are using regularization for our statistical model training—after fitting a gamma distribution—we transform the 2-month aggregated soil water data to a standard normal distribution (McKee et al. 1993), known as the standardized soil moisture index (SSI-2). Since SSI-2 is a proxy for the soil moisture levels, we simply refer to the precursor soil moisture (SM). The motivation for using SST is described in the introduction. Soil moisture was initially added to inform on the exposure to droughts occurring from April to July. However, results show that soil moisture at longer lead times adds value by “integrating exposure to (dominant) weather patterns” (see results and discussion sections 3 and 4).

For the U.S. soybean yields, we start with county-scale census data spanning from 1950 up to 2019 from the U.S. Department of Agriculture (USDA) National Agriculture Statistics Survey (NASS) Quick Stats database (http://www.nass.usda.gov/Quick_Stats, last accessed 1 March 2021). The dataset is first regridded to 0.5° spatial resolution. A grid cell contained within a county is assigned the yield value of that county. Otherwise, if several counties are contained within a grid cell, the grid cell is assigned the average yield value of the contained counties. In a second step, we select for grid cells with common sowing dates (i.e., mid-April–mid-May) and a soybean production area share of at least 90% rainfed agriculture. The period from April to May represents sowing dates for the majority of soybean production across the United States. Information on the soybean growing season dates and the production system used is obtained from the monthly irrigated and rainfed crop area database around the year 2000 (MIRCA2000), a global gridded dataset at a 0.5° resolution (Portmann et al. 2010).

b. Clustering of soybean production regions and derivation of spatial mean soy yield

Previous work showed that the mid- and southern regions are particularly sensitive to summer high maximum temperatures, while the northern regions are mainly sensitive to early summer cold minimum temperatures (Hamed et al. 2021). This north–south separation in weather sensitivity is ostensibly leading to a detectable difference in yield variability. We use a clustering algorithm to separate the northern soy producing regions from the mid- and southern regions (Fig. 2). For our gridded yield data, some grid cells span from 1950 to 2019, but not all. However, the clustering algorithm requires complete time series (no not-a-number). Hence, a trade-off exists between using fewer-but-longer observational time series versus more-but-shorter time series. We select data from 1975 onward since many regions cover the post-1975 period. We interpolate missing data using a second-order spline, extrapolation is done using a linear trend line. If more than 7 years of the time series needs to be extrapolated, the observational time series is excluded from the clustering analysis. Irrespective of the clustering method used (k-means, hierarchical agglomerative clustering optimizing intracluster correlation, or minimizing intracluster variability), we consistently found the same clusters when setting n clusters to 2. Sections 3a–3d focus on predicting poor yield events in the southern cluster 1, as this domain is coherently sensitive to weather (hot–dry) conditions (Hamed et al. 2021). We do not construct localized (county level) forecasts, as predicting at such a small spatial scale will decrease the signal-to-noise ratio. Such localized forecasts would require a more dedicated analyses and more in-depth knowledge to account for external factors that are specific for each county, although the signal-to-noise ratio issue will remain a challenge. Section 3e explores the skill on the U.S. state level.

We can now use all data (since 1950) and calculate an area-weighted mean time series of the observations that fall within cluster 1 or one of the U.S. states. The data are first detrended and standardized (Fig. A1; appendix A) at the gridcell level to focus exclusively on crop yield anomalies relative to the local expected value. In this way, the time-mean yield value at any grid cell is zero irrespective of varying yield potentials across the spatial domain. To get a sufficiently reliable estimation of the mean and standard deviation, we select data with a minimal length of 30 years. To simplify interpretability, we focus on relative poor harvest events within the spatial domain, that is, we do not apply weighting proportional to the area or the mean yield per grid cell. Poor harvest events are defined as years in which the spatial mean time series falls below the 33rd percentile threshold, that is, the 1-in-3 poor harvest years.

As mentioned in the introduction, low-frequency variability in the Pacific plays a role in the mechanism that leads to predictability in the eastern United States. Therefore, the detrending is done linearly, since any nonlinear detrending method will remove the small amount of slow variability in the crop time series that is in phase with the Pacific variability. Consequently, nonlinear detrending methods would to some extent disrupt the low-frequency covariability that exists.
between Pacific SST and crop yield fluctuations, and therefore disrupt the training of a statistical model.

c. Cross validation and preprocessing

The results presented in the main text are based upon the leave-three-out (LTO) (Iizumi et al. 2021) and the (operational-like) one-step-ahead (OSA) cross-validation (CV) techniques (Lehmann et al. 2020). LTO does not use the year prior and after the test year for training to reduce information leakage from adjacent years that could be present due to temporally correlated time series. This is repeated for each test year, meaning we have 69 training folds, with each 66 (or 67 at the edges) years of data available for training (Fig. A2). With one-step-ahead CV, we aim at emulating an operational-like setting. Thus, the crop yield preprocessing is done using only (and all available) training data from the past (detrending, standardizing, and calculating the event thresholds; see Fig. A4). For example, with the “one-step-ahead-25” CV, we forecast the recent 25 years (1995–2019). When predicting the year 1995, we only use data from 1950 to 1994. When predicting the year 1996, we only use data from 1950 to 1995, and so forth. Thus, the size of the training dataset varies between 44 and 68 years. Only

Fig. 1. Overview of data-driven pipeline(s). The preprocessing steps and input data are described in sections 2a–2c. We test four different feature input sets (sections 2d–2f) to predict the mid-to-southern U.S. cluster yield time series. The selected feature time series are also used to predict yield at the state level. Statistical models and verification metrics are described in sections 2g and 2h, respectively.

Fig. 2. (a) The clustering algorithm aims at minimizing the intracluster variability. The two detected regions are in line with previous results documenting the geographical differences in sensitivity to weather anomalies between northern and southern states. (b) The mid-to-southern cluster is sensitive to hot-dry weather, while the northern cluster is not (Hamed et al. 2021). Result sections 3a–3d only focus on predicting spatial averaged yield in cluster 1. Growing stages are adapted from Fig. S3 of Ortiz-Bobea et al. (2019).
finding the two spatial clusters of soy yield (section 2b) and the preprocessing of SST and SM precursor datasets (detrending, removing seasonal cycle, and calculating the SSI-2) is done in sample. We expect the latter effect to be small given the large trend and high variability in crop yield compared to SST and SM. Hence, this approach should give a good estimation of the operational forecast skill.

d. RGDR

In a response-guided dimensionality reduction (RGDR) method, the dimensionality reduction method takes into account the target variability (Bello et al. 2015). A successful and data-efficient approach is to compute pairwise correlations with a target time series (Kretschmer et al. 2017) and cluster spatially adjacent grid cells into precursor regions using via a clustering approach (Lehmann et al. 2020; Vijverberg et al. 2020). For the SST data, one-dimensional precursor time series are calculated for each cluster (here called precursor regions) by calculating an area-weighted and correlation-weighted spatial average (Fig. 3). For the soil moisture data, we calculate the spatial covariance time series of the correlation pattern, only considering significantly correlating grid cells (Fig. 4).

The clustering of the correlating grid cells is done via the density-based spatial clustering of applications with noise (DBSCAN) algorithm (Ester et al. 1996; Schubert et al. 2017). DBSCAN iteratively groups together points that are close, while separate and distant points are seen as outliers. With respect to a single point (i.e., grid cell), the radius parameter determines the distance at which points are grouped together. The distance is measured as the great-circle distance in km. If

FIG. 3. Schematic of the RGDR method applied to the preprocessed soy yield and SST data (subdomain shown), with the forecast issued the 1 Aug. Correlation maps are computed at four lags (here JJ, AM, FM, and DJ), only the first (JJ) and last (DJ) lag are shown. The significantly correlating grid cells ($\alpha_{FDR} = 0.05$) are clustered using DBSCAN. These clusters are used as masks to calculate area- and correlation-value-weighted region mean time series. All time series are standardized based on the training data.

FIG. 4. Schematic of the RGDR method applied to the preprocessed soy yield and SM data, with the forecast issued 1 Aug. Similar to Fig. 3, but here the spatial covariance time series of the correlation pattern is computed to create a single precursor time series.
there are fewer than 3 points in its vicinity, the significantly correlating grid cell is discarded as an outlier. Note, the clustering of positively and negatively correlating grid cell are treated separately. DBSCAN tends to create large clusters as the reachability of a cluster increases via the points at the edges, which can iteratively search for their nearby points. We avoid this by setting the radius parameter to a relatively low value, at 250 km. However, this can generate relatively nearby—but separate—clusters. For the model fitting and feature selection this is undesired since the strong spatial correlations can make these time series dependent, that is, carrying the same signal. Hence, after clustering with a low radius parameter, we search for closely located clusters and group those who show a strong intercorrelation ($\sim>0.4$). See appendix B for more information.

For the correlation maps, we set $\alpha = 0.05$ and account for multiple-hypothesis testing by applying the Benjamin–Hochberg correction (Benjamini and Hochberg 1995; Wilks 2006). The advantage of the RGDR is that the time series are tailored toward the target variable, while the detection power is high (not data hungry since initial step is based on correlation). The RGDR method can flexibly search for precursor time series at multiple lags (Vijverberg and Coumou 2022), thereby also considering the evolving spatial extent of the precursor regions (see e.g., Fig. 2). We will search for precursor time series at four lags, with the lag 1 being the 2-month mean of the 2 months prior to the forecast release date. The last lag (lag 4) is 8 months prior to the forecast release date. The added value of multiple-lag input is also benchmarked versus using only the most recent lag to forecast soy yield (see Table 3).

e. Causal inference as precursor selection method

Removing the spurious precursors from the input features reduces the risk of overfitting (Kretschmer et al. 2017). We rely on a causal inference approach to remove spurious precursor time series and thereby remove redundant information. The selection step is purposefully made not very strict as missing physical drivers is more detrimental to the forecast quality than having a few spurious ones. Via cross validation and hyperparameter tuning, we aim at assigning low weights to the small number of potential spurious features that pass the selection step (section 2g). To improve interpretability, we use one simple rule: the time series of a precursor region at any given lag should always be dependent given the influence of every other precursor detected at any lag. This way we will not regress out the influence of autocorrelation as this information is needed to learn the evolution of a precursor region that can enhance predictability (Switanek et al. 2020b; Vijverberg and Coumou 2022). We use partial correlation for our conditional independence tests (Ebert-Uphoff and Deng 2012). Although we do not rely on more sophisticated causal discovery (Runge et al. 2019), the causal inference step is expected to keep the strongest correlating precursor time series and filter out (most) time series that are correlating due to a common driver effect or an indirect link. Due to this simplification, we refer to the selected precursors not as causal but as conditionally dependent, that is, significantly correlated with the target variable even when conditioned on each time series in vector $Z$ (Fig. 5). (The result of this selection step is visualized in Figs. 6 and 7.)

f. Baseline dimensionality reduction approach

We compare our rather complex dimensionality reduction approach with a simple benchmark. We compute, for each training dataset, the first empirical orthogonal function (EOF) over the PDO domain ($110^\circ–260^\circ$E, $20^\circ–70^\circ$N) from our preprocessed SST dataset using the November to March months, closely resembling the PDO index (Trenberth and Fasullo 2013). By projecting the EOF loading pattern on observations, we extrapolate to the test sets. We also compute the area-weighted spatial mean time series over the ENSO-3.4 domain ($5^\circ$S–$5^\circ$N, $190^\circ–240^\circ$E) as a proxy for ENSO variability. We refer to this baseline approach as “climate indices.”

g. Statistical models and hyperparameter tuning

We tested both a regularized logistic regression (LR) and random forest (RF) to make probabilistic forecasts. For the tuning of hyperparameters, we apply a double cross-validation approach (Vijverberg et al. 2020). In this approach, each training dataset of the “outer” cross validation is split into training and validation sets using a second “inner” cross-validation. The methods for the “outer” cross-validation are introduced in section 2c, the “inner” cross validation is always a tenfold CV. A schematic of the double cross validation approach is shown in Fig. A2. The parameters are tuned to minimize the Brier score (BS). The statistical models, cross-validation and parameter-search software are used from the Scikit-learn Python package (Varoquaux et al. 2015). To ensure equal penalty of the regularization parameter $C$ (Table 1), all features are standardized (based on the training data) prior to model fitting. For the random forest, we tune the max depth, indicating the depth of each tree (number of splits). With max features, we limit the

---

**Selection algorithm**

\[ X = (x_1^1, x_2^1, ..., x_n^1) \]

\[ X: \text{set of all RGDR features} \]

for $x_i^1$ in $X$:

\[ Z = X \setminus \{x_1^1, ..., x_i^1\} \]

\[ Z: \text{set of all features, excl. } x_i^1 \text{ and lag shifted timeseries of } x_i^1 \]

list_pvals = [ ]

for $z$ in $Z$:

\[ p_{val} = \text{parcorr}(x_i^1, y | z) \]

\[ \text{list_pvals.append}(p_{val}) \]

if max(list_pvals) < 0.05:

\[ x_i^1 \text{ was always conditionally dependent given } z \]

\[ x_i^1 \in S \]

**S: set of Conditionally Dependent features**

---

FIG. 5. Pseudocode of selection algorithm. Where “parcorr” indicates a partial correlation analysis, $x_i^1$ is the to-be-tested time series, $y$ is the soy yield time series, $z$ is the to-be-tested confounding time series, and $p_{val}$ is the $p$ value.
percentage of features used to create a single tree to 40% or 80%. With max samples, we limit the percentage of samples used to construct each tree to 40% or 70%. Lower percentages of the latter two parameters can improve generalizability. The N estimators refer to the number of trees built.

### h. Skill metrics

Multiple metrics are needed for proper verification (Vijverberg et al. 2020; Wilks 2011) and we use 1) the Brier skill score (BSS), where the benchmark is the observed climatological probability after the trend line has been subtracted (i.e., 33% probability of poor harvest), 2) the accuracy, and 3) the precision metric (Table 2). The probability threshold used for the latter two metrics is equal to the climatological probability (33%). Given the probability of poor harvest, the accuracy and precision of a random guess is 54 and 33, respectively.

### 3. Results

The results of the feature selection algorithm, using the LTO CV, are discussed in section 3a. The corresponding figures for the OSA CV can be found in appendix B. The forecast skill of LTO and OSA are presented in sections 3b and 3c, respectively. Section 3d shows an overview of the forecast skill when using different input features, different CV methods and the two statistical models. The forecast skill on a smaller spatial domain (state level) is verified in section 3e.

#### a. Conditionally dependent (CD) precursor regions

The Pacific horseshoe-like region is the most robust SST precursor region detected at all lead times (Fig. 6), even up to 15 months prior to the start of the harvest period (1 October). The spatial extent of the horseshoe region decreases as function of lead time, but the magnitude remains strong. We also observe a robust Atlantic signal in summer and late spring (i.e., at short lead times), which is interpreted as the SST response to the westward extension of Rossby wave (high and low pressure system over the Atlantic) that is associated with the low pressure system (linked to less hot-dry conditions) over the eastern United States. However, the fact that the Atlantic regions passes all conditional independence tests (in all training samples) suggests that the western Atlantic SST variability is not solely the result of the Rossby wave that is forced by the horseshoe Pacific pattern (see discussion section 4).

The positive soil moisture correlation pattern in August shows that locally higher soil moisture levels correlate with higher end-of-year yields. For the northwest of North America (i.e., outside the harvest area), we observe regions with negative correlations (Fig. 7) indicating wet soils in the northwest are linked to poorer harvests in our target area (cluster 1). The correlation maps prior to August no longer show a soil moisture signal within the harvest area, and they show a positive correlation over the northwest of North America domain (opposite sign compared to August). The circulation patterns associated with the soil moisture correlation patterns are shown in Fig. B1. We find that the soil moisture pattern correlates with a circulation pattern that shows a consistent low pressure system over the mid-Pacific and high pressure system over the North/northeastern Pacific. Via ocean–atmosphere interaction, such a circulation pattern is likely able to strengthen the Pacific horseshoe pattern (see discussion section 4). At very long lead times, this indirect signal becomes weaker, and the soil moisture is generally filtered out by the conditional independence tests.

#### b. LTO hindcast skill

Using the random forest model, the out-of-sample hindcast soy harvest failures between 1950 and 2019 can be predicted with good accuracy already in February, which is approximately 3 months prior to sowing (Fig. 8). This long-lead predictability is possible during specific windows of predictability. We use the state of the horseshoe-like Pacific precursor pattern (Fig. 6) to quantify the signal strength, which is used to identify the windows of predictability (i.e., when the signal is strong). For a given year, the state of the horseshoe precursor pattern is calculated by taking the mean over all lags that passes the conditional independence tests. We use both the horseshoe Pacific and the soil moisture conditions for the August forecast, since we know that local end-of-summer SM can strongly impact crop growth (Hamed et al. 2021). Using soil moisture earlier in the season to identify the window of predictability did not improve results. The signal strength (S) time series are plotted below each forecast in Fig. 8. The years with an

### Table 1. Exhaustive search over hyperparameters is tested within a tenfold CV aiming to minimize the BS (error score).

<table>
<thead>
<tr>
<th>Statistical model</th>
<th>Hyperparameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 regularized LR</td>
<td>C (inverse regularization parameter) [1 \times 10^{-2}, 1 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}, 0.5, 1, 1.2, 4, 7, 10, 20]</td>
</tr>
<tr>
<td>Random forest classifier</td>
<td>Max depth [2, 5, 8, 15]</td>
</tr>
<tr>
<td>(n estimators = 300)</td>
<td>Max features [4, 0.8]</td>
</tr>
<tr>
<td></td>
<td>Max samples [4, 0.7]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistical model</th>
<th>Hyperparameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 regularized LR</td>
<td>C (inverse regularization parameter) [1 \times 10^{-2}, 1 \times 10^{-2}, 5 \times 10^{-2}, 1 \times 10^{-1}, 0.5, 1, 1.2, 4, 7, 10, 20]</td>
</tr>
<tr>
<td>Random forest classifier</td>
<td>Max depth [2, 5, 8, 15]</td>
</tr>
<tr>
<td>(n estimators = 300)</td>
<td>Max features [4, 0.8]</td>
</tr>
<tr>
<td></td>
<td>Max samples [4, 0.7]</td>
</tr>
</tbody>
</table>

### Table 2. Verification metrics presented in this manuscript. The first row shows the BSS with the climatological probability as the benchmark forecast (0.33). The BS is calculated for both the benchmark (BS_{clim}) as well as the forecast (BS_f). The term \( f_i \) represents the forecast at time step \( i \), and \( N \) is the number of forecast/observation pairs.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSS</td>
<td>( BS_f = \frac{1}{N} \sum (f_i - o_i)^2 )</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Accuracy</td>
<td>( BS_f = \frac{1}{N} \sum (f_i - o_i)^2 )</td>
<td>Precision</td>
</tr>
<tr>
<td>Precision</td>
<td>( BS_f = \frac{1}{N} \sum (f_i - o_i)^2 )</td>
<td>True positive + true negative</td>
</tr>
<tr>
<td>Precision</td>
<td>( BS_f = \frac{1}{N} \sum (f_i - o_i)^2 )</td>
<td>True positive</td>
</tr>
<tr>
<td>Precision</td>
<td>( BS_f = \frac{1}{N} \sum (f_i - o_i)^2 )</td>
<td>True positive + false positive</td>
</tr>
</tbody>
</table>

**BS_{clim}** is the number of forecast/observation pairs.
anomalously strong state, either negative or positive, are indicated by red and blue dots, respectively (see legend in Fig. 8). During these states, we expect larger deviations from the climatological mean weather, and therefore better yield predictability.

The tables next to each forecast in Fig. 8 show three skill metrics, that is, the BSS, the accuracy, and the precision (see methods in section 2h). The columns refer to the subset of years that are included when calculating the verification skill: All (all years), top 50% (the 50% years with strongest signal), and top 30% (the 30% years with strongest signal). Hence, these subsets of years indicate the windows of predictability in time. During the top 30% years, that is, when the signal is very strong, a high Brier skill score (>0.5) shows that the forecasts are both reliable and confident in terms of its assigned probability. Overall, we observe a decline in skill as function of lead time. However, during the top 30% window of predictability the

---

**Fig. 6.** SST (2-month mean) correlation maps vs the crop yield variability in cluster 1 (see Fig. 2) for each forecast month. A correlation value is only shown if a grid cell is significantly correlating in one of the 69 training datasets. The green integers denote the number of training samples the precursors time series has passed all the conditional independence tests. The blue integers denote the number of training samples the precursor time series is detected by the RGDR. For clarity, we only show the regions that were conditionally dependent in at least 50% of the training samples. The precursor region labels assigned by DBSCAN are shown in Fig. B2.
skill for poor harvest years (<33rd percentile) is remarkably high across lead times, with the August forecast achieves a BSS of 0.80 and precision of 88% and the February forecast achieves a BSS of 0.59 and precision of 62%. We obtain substantially less forecast skill when predicting good harvest years (>66th percentile), result not shown (see discussion in section 4).

c. One-step-ahead forecast skill

In an operational-like setting, our forecast system would have been able to predict low yield years with high skill over the last 25 years (Fig. 9). Again, we observe a systematic boost in forecast skill during the windows of predictability. The top 50% forecast (consisting of 14 years) in February achieves a precision of 67%, that is, 4 true positives out of 6 forecasted low-yield events (see Fig. 9). Out of the total 15 predictions, 12 were correct, that is, an accuracy of 80%. The precursors that are selected are similar to precursors presented in section 3a, although the horseshoe Pacific SST region is more often removed after conditioning on the soil moisture pattern time series for the August and June forecasts. For the OSA cross-validation, we are more limited in the amount of data points for training and verification. Calculating the skill metrics based on only 25 (or fewer) forecast/observation pairs can introduce a (sampling) bias by being (un)lucky in the sample set. To get a feeling for the sampling bias, we also calculated the skill metrics over the recent 30 and 20 years (Fig. C1) and we slightly changed the event time series by changing the quantile threshold from 0.31 up to 0.35 with steps of 0.01 (Fig. C2). Given these perturbations, the precision did not drop below 65% for the February forecast during the window of opportunity (top 50%). Overall, the skill is robust across months and quantiles, yet there are some (positive and negative) outliers (Figs. C1 and C2). We also tested the influence of a varying event frequency within the verification periods and found this to be minor (appendix C).

d. Synthesis of results

Within our set of experiments shown in Table 3, using combined lead time models (features at multiple lags) show the largest boost in forecast skill compared to using only the most recent lag (indicated by “lag 1 RGDR precursors” in Table 3). Furthermore, we observe that, with enough data (66 years),
the random forest performs best when using all precursor time series extracted by the RGDR method, that is, no feature selection (vector \( \mathbf{X} \) in Fig. 5). With limited training data (44 up to 68 years) the performance of the regularized LR and the RF is very similar. The RF tends to be a bit more careful in its assigned probabilities (lower sharpness) and we believe this causes it to perform a bit more stable. For example, we tested the impact of in-sample versus out-of-sample preprocessing of the target variable, and the RF was less sensitive to this change (Fig. C2).

e. State-level forecast skill

When predicting the state-level yield (Fig. 10) forecast skill is good (BSS of \( \geq 0.3 \)) inlands of our mid-to-southern target cluster (section 2b), yet underperforms compared to predicting the average of the entire target cluster (Fig. 9). The state-level forecasts are made using the one-step-ahead-25 cross validation and the random forest (similar skill was obtained using the regularized logistic regression). We use only the selected—more trustworthy—features and refitted the model for each state.

4. Discussion

Our forecast framework provides new opportunities for stakeholders to make better-informed decisions already early in the season. When a poor harvest (defined as a 1-in-3-yr event) is predicted on 1 February during a window of opportunity (top 50%), that is, \( \sim 3 \) months before sowing, the risk of a poor harvest increases from 33\% to 65\% (Fig. 9). Using the window of predictability concept, we can communicate when we have high confidence in a forecast and when not.
Based on a forecasted poor harvest in February, farmers still have sufficient time to change their planting timing, avoid planting the drought-prone agricultural fields, or select more drought/heat-resistant soy cultivars. Such forecasts are also relevant for (non)governmental institutions, commodity traders and crop insurance companies (Basso and Liu 2019; Torreggiani et al. 2018). Forecasts based on surveys cannot provide farmers with this information at sufficient lead times (Beguería and Maneta 2020; National Agricultural Statistics Service 2012). To the best of our knowledge, current long-range dynamical weather forecasts are also unable to inform farmers on the weather-related risks at these very long lead times. Although a comparison is not within the scope of this paper, the North American Multimodel Ensemble reforecast (1982–2009) shows negative skill for the JJA temperature in the eastern United States when initialized 1 January (Kirtman et al. 2014). Similarly, the European fifth-generation dynamical seasonal forecast model (SEAS5) shows relatively low correlation coefficients (roughly 0.4) for summer temperature in our target region (Johnson et al. 2019). Another benefit of data-driven forecasts is that predictions of impact (e.g., poor harvest years) are as straightforward as predicting weather variability (e.g., hot-dry extremes), whereas predicting poor harvest with plant simulation models is very difficult (Brown et al. 2018; Iizumi et al. 2018). For data-driven predictions, a proper verification is crucial, which is why we have used strict train-test splitting, multiple metrics (importance illustrated in Vijverberg et al. 2020), and a clear benchmark forecast after subtracting the linear trend. This is unfortunately still a major issue in statistical crop forecasting (Schauberger et al. 2020).

To achieve the high forecast skill, we found that the following steps are important: 1) clustering of target variable, 2) using the horseshoe Pacific region (extracted by the RGDR method) to identify the windows of predictability, and 3) using multiple lags as input. First, the spatial clustering algorithm identifies spatial regions that behaved similarly in terms of harvests. The southern cluster (cluster 1 in Fig. 2) encompasses the mid-to-southern producing region, which is sensitive to hot-dry extremes in summer, while the northern region is less sensitive to weather (Hamed et al. 2021; Schauberger et al. 2017a). We

![Fig. 9](image-url)
also obtain no forecast skill for the northern cluster 2 (result not shown). For the mid-to-southern region we obtain positive yet substantially less forecast skill for positive (>66th percentile) yield events (result not shown). This is expected given the strong nonlinear temperature response of crop growth, that is, temperatures exceeding 30°C are very harmful, while the sensitivity to temperature below the 30°C mark is much weaker (Schlenker and Roberts 2009; Schauberger et al. 2017b). To summarize, optimizing the signal-to-noise ratio of the target crop time series is a crucial step to improve skill on the subseasonal-to-seasonal time scales, as has also been shown for temperature forecasts (Vijverberg et al. 2020). The second step, that is, identifying the window of predictability based on the state of the Pacific, also resulted in a systematic and substantial boost in skill, in line with previous results (Vijverberg and Coumou 2022). As shown in Table 3, forecast skill is substantially increased by including the evolution of precursor patterns throughout the season, rather than just using a snapshot of their state at forecast time. Forecasting yield at the state level reduces skill, which is expected as the effects of unobserved factors become more dominant (such as crop management decisions, observational biases, or for example, unpredictable local deep convection events) on a smaller spatial scale. Nevertheless, the state-level predictions are still skillful in 6 states, which are responsible for ~31% of total soy production.

The reduced set of precursors by applying conditional independence testing (i.e., ∼10 precursors instead of ∼40) enables a better physical interpretation (see section physical interpretation below). Moreover, this precursor selection step reduces the risk of overfitting. Figure B3 shows the precursor regions that were removed by the selection step. Based upon expert judgement, these regions (often small and located close to the coastline) are likely not causally linked to our target. Hence, although the selection step can slightly degrade skill metrics (Table 3), the trustworthiness is increased. However, we did notice that for the OSA-25 CV (with less training data), the selection step was less robust.

In general, there are still opportunities to further improve forecast skill. Here we used DBSCAN to cluster coherent precursor regions. We assume that grid cells that 1) correlate with the same sign and 2) are located close to each other carry the same signal. The spatial mean that we calculate for each precursor region reduces noise, thereby improving the SNR, yet particularly for large precursor regions, there are likely spatial differences in the signal strength. For this reason, the spatial mean is weighted with the correlation value, but one might want to model this with more detail. Furthermore, since crop yield is also affected by other factors than weather, it will also be insightful to directly predict hot–dry conditions. In addition, the selection step can be further fine-tuned by relying more on expert knowledge.

**Physical interpretation**

Our dimensionality reduction and precursor selection method based on conditional independence tests has identified the horseshoe Pacific SST region and the soil moisture

---

**FIG. 10. State-level forecast skill in terms of the BSS.** The data-driven pipeline is similar to Fig. 9, yet here we forecast the poor (1–3) yield events based on the out-of-sample preprocessed state-level aggregated yield. Forecasts are made for April, March, and February.
patterns as the most robust precursors. The very long-lead-time signal exists due to the low-frequency dynamics in the Pacific, which is often summarized using the PDO index, although the PDO is not driven by a single mechanism (Newman et al. 2016). The persistence of the (horseshoe Pacific) SST anomaly is crucial to force a barotropic Rossby wavelike response, as supported by modeling experiments (Ferreira and Frankignoul 2005) and observations (Vijverberg and Coumou 2022). We argue that this is the physical reason behind the increased predictive skill when using multilagged input features, that is, having information on multiple lags informs on the persistence and momentum of the SST precursor. Strong signal years are therefore also characterized with a high persistence by calculating mean over all lags and subsequently selecting anomalous states.

Besides the local effect of soil moisture on crop yield (Hamed et al. 2021), we found that the soil moisture pattern over the continental North America reflects the dominant circulation patterns that are present prior to the summer season. The circulation associated with the soil moisture patterns is, based on previous research, expected to strengthen the horseshoe Pacific SSTs (Vijverberg and Coumou 2022). Causal discovery analyses on observations showed that the slowly varying horseshoe Pacific region promotes the occurrence of this arcing Rossby wave and that it is also strengthened by it. Both low-frequency ocean dynamics in the Pacific and atmosphere-ocean forcing (the latter captured indirectly via soil moisture) determine the development of a strong ocean-atmosphere boundary forcing. Soil moisture has a much longer memory making it better suited for forecasting purposes, that is, it is much less noisy compared to using circulation directly. In June–July, the circulation pattern (Fig. B1) also projects onto the concomitant robust SST Atlantic signal. However, the fact that the Atlantic SST signal is not conditionally independent when regressing out the influence of the horseshoe Pacific pattern, suggests another cause could be present. We suspect that dominant periods of the wavenumber 6 Rossby wave, which is a known summer mode of variability (Branstator and Teng 2017; Kornhuber et al. 2017; Vijverberg and Coumou 2022), also plays a role in driving both yield variability as well as Atlantic SST variability.

5. Conclusions

We show that a good physical understanding in combination with innovative data-driven techniques that aim at optimizing the signal-to-noise ratio can achieve high forecast skill at unprecedented lead times. Not all spatial regions and years have the same level of intrinsic predictability (Mariotti et al. 2020). To detect regions and periods with high predictability (i.e., windows of predictability), both target and input features need to be carefully selected and optimized (Table 3), see also (Vijverberg et al. 2020; Vijverberg and Coumou 2022). To do so, we apply a clustering technique for the target, and we use a response-guided dimensionality reduction method and a feature selection based on causal inference. We target the 1-in-3 poor soy harvest years within an aggregated spatial domain (Fig. 2) showing a homogeneous sensitivity to hot–dry extremes (Hamed et al. 2021). Our operational-like forecast system can predict poor harvest years with high forecast skill and high confidence already on the 1 February, that is, 3 months prior to sowing (Fig. 9).
This forecast can be released eight months prior to the harvest period, whereas the current operational forecast system, although released on a higher spatial resolution, is released only in August (Schnepf 2017). If we forecast a poor harvest, during a year with an anomalous horseshoe Pacific SST state, the probability of a poor harvest released in February increases from the normal 33% to 65% (Fig. 9). The high Brier skill score (>0.5) shows that the forecast is both reliable and confident in terms of its assigned probability. Most importantly, our forecast is released 3 months prior to sowing, which allows farmers to take anticipatory action, that is, change to more drought resistant cultivars or change planting management. Our approach can be tuned to specific needs of stakeholders, for example, focus on specific subregions, adapting the threshold that defines a poor harvest year, and making additional forecasts for hot–dry weather to better isolate the weather-induced risk.

Acknowledgments. This research was supported by the Dutch Research Council under the grant agreement 016.Vidi.171011 (Vidi project: Persistent Summer Extremes), by the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement 820712 (project Remote Climate Effects and Their Impact on European Sustainability, Policy and Trade (RECEIPT]) and by the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement 101003469 (project XAIDA, eXtreme events: Artificial Intelligence for Detection and Attribution).

APPENDIX A

Preprocessing of Crop Yield Data

After the clustering (as described in the method section 2) has been performed, we aggregate all data that is located within the southern cluster (label 1). We assume (as supported by literature), that the crop growth in each grid cell is negatively influenced by hot–dry conditions. However, there are clear differences within the cluster in terms of absolute productivity, interannual variability, long-term trend, and time period that the observations cover (Fig. A1). Therefore, we first detrend (Fig. A1a) and standardize the time series (Fig. A1b), before calculating the spatial mean. Figure A1b also shows that prior to ~1975 the variability is relatively small (most grid cells vary between $-1\sigma$ and $+1\sigma$), while during the 1980–2019 period most grid cells vary between $-3\sigma$ and $+2\sigma$. The outer cross-validations are introduced in section 2c. For the tuning of hyperparameters, we apply another “inner” cross-validation scheme (Vijverberg et al. 2020), meaning that each “outer” training dataset is split into (inner) training and validation sets using a tenfold CV approach. A schematic of the double cross-validation approach is shown in Fig. A2. Figure A3 shows the shift of producing regions toward lower latitudes around the 1970s. The out-of-sample preprocessing is visualized in Fig. A4.

Data availability statement. Table 4 shows the links to the publicly available data sources. The Python code and gridded USDA soy yield data (as described in section 2a) are available at https://doi.org/10.5281/zenodo.7498927.
FIG. A1. (a) For 40 grid cells, time series of crop yield (bushels ha\(^{-1}\)) (black line) and the fitted linear trend line (red line). The 40 grid cells are roughly evenly distributed over the mid-to-southern cluster. (b) The 40-gridcell time series of crop yield (red lines) after detrending and standardizing. The black line shows the mean over all time series within the mid-to-southern cluster that contain more than 30 datapoints.
Outer Leave-Three-Out CV, example of second split:

Input target dataset (69 years)

1951 1952 1953 1954 ... 2016 2017 2018 2019

Dimensionality reduction and feature selection (66 years)


Inner random 10-fold CV, example of single split:

Tune hyperparameters

1954 1955 ... 1979 ... 1990 ... 2002 ... 2008 ... 2011 ... 2018 2019

= train  = validation  = test  = not used

Fig. A2. Schematic of double cross validation, showing the second split of the leave-three-out (LTO) outer CV. Note that we only predict a single year out of sample per training fold. Hence, this process is repeated until all years are predicted (i.e., there are 69 outer training folds). The inner CV is always a random tenfold CV.

Fig. A3. Coverage of datapoints in percentages per time period of the mid-to-southern cluster (label 1 in Fig. 2). The “total # of years” refers to the number of years within each time period (23, 23, and 24 years).
FIG. A4. As in Fig. A1, but here done out-of-sample for the one-step-ahead-25 cross validation. Blue indicates the first training dataset \((n = 44)\), red indicates the last training dataset \((n = 68)\). The trendline and standardized time series are here extrapolated to all remaining datapoints, yet in our final out-of-sample preprocessed time series, the extrapolation of the trend and standardization is done for a single year for each training dataset.
APPENDIX B

Response-Guided Dimensionality Reduction and Causal Precursor Selection

As discussed in section 3a, Fig. B1 shows that the soil moisture correlation patterns are associated with dominant circulation patterns that are known to strengthen the horseshoe Pacific SST state. More anomalous states (positive or negative) of the horseshoe Pacific are associated with stronger boundary forcing for the atmosphere, and therefore a higher signal-to-noise ratio (Vijverberg and Coumou 2022).

Figure B2 shows the precursor regions as identified by the RGDR method for the LTO cross validation. Figure B3 shows the SST precursor regions that were filtered out by the precursor selection step after being identified by the RGDR method for the LTO cross validation. Note that the spuriously correlating regions are often located close to the coast. Intuitively, SST variability close to the coast is much more affected by local small-scale dynamics. Due to the lack of large-scale spatial correlation, there are more independent realizations of SST variability that increase the probability that a time series strongly correlates by chance. Due to the high (spurious) significance, the Benjamin–Hochberg false discovery rate correction that is applied for the correlation maps is not sufficient to filter these out.

Figures B4 and B5 show the robust SST regions and soil moisture patterns for the one-step-ahead cross validation as discussed in the results section 3c.

In DBSCAN, the radius (eps parameter) of 250 is chosen to define neighboring grid cells, which is found to produce regions of reasonable sizes and spatial separation. Well separated very small regions (size of ~1 grid cell) regions are automatically ignored. As explained in the method section 2, DBSCAN tends to create one single very large precursor region (the horseshoe Pacific region), while adjacent smaller regions are kept separate. This could lead to physically nonsensible partial correlation tests since adjacent regions are expected to correlate due to their spatial proximity. Hence, in a second step we find precursor regions that are close to each other, but before clustering them together, we validate if they are indeed sufficiently correlating (coefficient should be approximately >0.4). In this second step, we calculate the intercluster haversine distance based on the center of each precursor region. Because we now use only the center latitude longitude location of the regions, the distances between the features are much larger compared to the first DBSCAN step (where grid cells were often right next to each other). Therefore, the radius (eps parameter) is now set to 2000, which clusters the regions together at a realistic distance, that is, they are expected to be correlating due to their spatial proximity.

Fig. B1. (top) The correlation maps of soil moisture vs end-of-year crop yield at different lags. The soil moisture correlation maps correspond to the first column of Fig. 7. (bottom) The correlation between geopotential height at 500 hPa vs the spatial covariance time series of the soil moisture patterns (mean time series over the training samples of the LTO cross validation).
FIG. B2. SST precursor regions found by the DBSCAN clustering algorithm. Each label is used as a mask to calculate area-weighted and correlation-weighted time series.
FIG. B3. SST precursor regions that were identified by the RGDR method with the leave-three-out cross validation, but were filtered out by the precursor selection step, i.e., they were found conditionally independent given the influence of another precursor region time series. The integers denote the amount of training datasets in which the correlating region was extracted by the RGDR method.
FIG. B4. SST (2-month mean) correlation maps vs the crop yield variability in cluster 1 (see Fig. 2) for each forecast month after the selection step. A correlation value is only shown if a grid cell is significantly correlating in one of the 25 training datasets. The green integers denote the number of training samples the precursors time series has passed all the conditional independence tests. The blue integers denote the number of training datasets the precursor time series is detected by the RGDR. For clarity, we only show the regions that were conditionally dependent in at least 13 of the 25 training samples. As in Fig. 6, but using the one-step-ahead 25-yr cross validation.
APPENDIX C

Forecast Verification

The observed low yield events (Fig. 8) show some multi-year variability in the frequency of events, with fewer events occurring between 1955 and 1974 and between 2013 and 2019 [16 events (10 yr)], yet more events happening between 1975 and 2012 [47 events (10 yr)]. This decadal variability can be expected, given the importance of the extratropical Pacific SST variability in affecting the weather in eastern United States. The Pacific is well known for its decadal variability associated with the Pacific decadal oscillation (Newman et al. 2016). Due to this decadal variability, the frequency of events (from here on called base rate) was 40% in the recent 25 years. This deteriorates the benchmark forecast skill and thereby could lead to “spurious” skill. However, changing the climatological benchmark to the true base rate in the test set had a negligible effect on the BSS. On the other hand, the forecast model is challenged as it has to operate in a climate where the base rate is different from what it was trained upon, the later being 33%. To further investigate, we apply the one-step-ahead CV and concomitant preprocessing over different time spans [1990–2019 (30 years), 1995–2019 (25 years), and 2000–2019 (20 years)], with base rates of 36%, 30%, and 36%. In general, we observe that the differences in skill during the window of predictability for the one-step-ahead 30, 25, and 20 fall within an expected sampling bias, with perhaps the one-step-ahead 30 performing slightly worse (Fig. C1). The relatively small amount of training data (38 up to 43 years) for the first 5 years could have played a role in the small drop in skill. Figure C1 shows that the high forecast skill is robust against changing the verification period. This suggests that the forecast generalizes well, even when confronted with different base rates in the test sets.

Figure C2 shows that the random forest model is less sensitive to overfitting, as the difference in skill between the out-of-sample preprocessing and in-sample preprocessing is smaller than for the regularized logistic regression. Given that the perturbations to the events are small, Fig. C2 also gives a hint for the impact of the sampling bias.

Fig. B5. SM (SSI-2) correlation maps vs the crop yield variability in cluster 1 (see Fig. 2) for each forecast month after the selection step. A correlation value is only shown if a grid cell is significantly correlating in one of the 69 training datasets. The SM precursor time series is based upon the spatial covariance of the (significant) correlation values. The ratio shows the conditional dependent/detected precursor time series, similar to Fig. B4. If the SM time series is not conditionally independent in at least 13 of the 25 training samples, the SM correlation pattern is completely masked. The spatial domain of cluster 1 is shown in light pink.
FIG. C1. One-step-ahead forecast skill validated over different periods (the recent 30, 25, and 20 years) for the poor yield events in the mid-to-southern U.S. cluster.

FIG. C2. Visualizing the impact of in-sample versus out-of-sample preprocessing of the target variables (mid-to-southern U.S. cluster) for the BSS. Using the one-step-ahead cross-validation over the last 25 years. In addition, we test the skill when using different quantiles to observe the impact of minor changes in the binary event time series.
REFERENCES


Kornhuber, K., V. Petoukhov, D. Karoly, S. Petri, S. Rahmstorf, and D. Coumou, 2017: Summertime planetary wave


