Cross-Validation Strategy Impacts the Performance and Interpretation of Machine Learning Models

LILY-BELLE SWEET,[a] CHRISTOPH MÜLLER,[b] MOHIT ANAND,[a] AND JAKOB ZSCHEISCHLER[a,c]

[a] *Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research—UFZ, Leipzig, Germany*
[b] *Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Germany*
[c] *Technische Universität Dresden, Dresden, Germany*

(Manuscript received 24 March 2023, in final form 31 May 2023, accepted 3 July 2023)

ABSTRACT: Machine learning algorithms are able to capture complex, nonlinear, interacting relationships and are increasingly used to predict agricultural yield variability at regional and national scales. Using explainable artificial intelligence (XAI) methods applied to such algorithms may enable better scientific understanding of drivers of yield variability. However, XAI methods may provide misleading results when applied to spatiotemporal correlated datasets. In this study, machine learning models are trained to predict simulated crop yield from climate indices, and the impact of cross-validation strategy on the interpretation and performance of the resulting models is assessed. Using data from a process-based crop model allows us to then comment on the plausibility of the "explanations" provided by XAI methods. Our results show that the choice of evaluation strategy has an impact on (i) interpretations of the model and (ii) model skill on held-out years and regions, after the evaluation strategy is used for hyperparameter tuning and feature selection. We find that use of a cross-validation strategy based on clustering in feature space achieves the most plausible interpretations as well as the best model performance on held-out years and regions. Our results provide the first steps toward identifying domain-specific "best practices" for the use of XAI tools on spatiotemporal agricultural or climatic data.

SIGNIFICANCE STATEMENT: "Explainable" or "interpretable" machine learning (XAI) methods have been increasingly used in scientific research to study complex relationships between climatic and biogeoscientific variables (such as crop yield). However, these methods can return contradictory, implausible, or ambiguous results. In this study, we train machine learning models to predict maize yield anomalies and vary the model evaluation method used. We find that the evaluation (cross validation) method used has an effect on model interpretation results and on the skill of resulting models in held-out years and regions. These results have implications for the methodological design of studies that aim to use XAI tools to identify drivers of, for example, crop yield variability.

KEYWORDS: Agriculture; Crop growth; Artificial intelligence; Machine learning; Model interpretation and visualization

1. Introduction

The changing climate has already affected agricultural systems worldwide (Iizumi and Ramankutty 2016; Brás et al. 2021), and projections of continuing warming, increasing heat extremes, and intensifying short-term precipitation events suggest that further impacts can be expected in the coming decades (Seneviratne et al. 2021). According to state-of-the-art crop model simulations, severe shifts in agricultural productivity due to climate change may occur within the next 20 years in several regions (Jägermeyr et al. 2021). Accurate assessments of future climate impacts on crop yields are essential for adaptation planning as the worldwide population continues to grow, but projections of global yields remain highly uncertain (Jägermeyr et al. 2021; Müller et al. 2021). This can partly be attributed to uncertainty in climate model projections, particularly in more data-scarce regions, but a large portion of the range in yield projections arises from variance between crop models (Ruane et al. 2021; Müller et al. 2021; Jägermeyr et al. 2021).

Recent studies have found that process-based crop models fail to capture the impact of extreme heat or drought events on yields (Heinicke et al. 2022; Lafferty et al. 2021). Furthermore, due to the complex and nonlinear relationships between weather, cultivar, management, soil, pests, diseases, and end-of-season crop yields (Lesk et al. 2022), the compounding of climate events that are not themselves extreme may result in extreme yield losses (Zscheischler et al. 2020; Wiel et al. 2020). An example of such an extreme impact from nonextreme climate events is the 2016 record-breaking wheat losses in France, which were unanticipated by forecasters (Ben-Ari et al. 2018). Additionally, the impacts of extreme events may be moderated by compounding conditions; for example, heat impacts on crops can be alleviated by wet conditions, while extreme wet conditions are linked to decreased yields in both warm and cool conditions (Hamed et al. 2021).

The increasing availability of agricultural data at multiple spatial scales has precipitated the use of data-driven methods to disentangle the complex interactions between climate and

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/AIES-D-23-0026.s1.

Corresponding author: Lily-belle Sweet, lily-belle.sweet@ufz.de

DOI: 10.1175/AIES-D-23-0026.1 e230026

crop yield. Statistical modeling has been used extensively for yield prediction, often using annual or seasonal aggregates of climatic variables such as temperature and precipitation (Ortiz-Bobea et al. 2019; Laudien et al. 2020; Ribeiro et al. 2020). Recent studies have included extreme indicators so as to capture the effect of short-term climate extremes (Vogel et al. 2019), and the use of machine learning models has been explored (Vogel et al. 2021). A growing body of research has shown the ability of such models to accurately predict agricultural yield (Crane-Droesch 2018; Liu et al. 2022; Mateo-Sanchis et al. 2023). However, the “black box” nature of these tools impedes the extraction of relationships from trained models for scientific study.

“Explainable” or “interpretable” artificial intelligence (XAI) methods purport to give insight into the relationships captured by machine learning models (Rudin 2019; Ryo 2022). These methods have been used to identify drivers of yield variability or yield failure events in the United States and Europe (Goulart et al. 2021; Peichl et al. 2021; Webber et al. 2020; Wolanin et al. 2020; Mateo-Sanchis et al. 2021; Martínez-Ferrer et al. 2020; Newman and Furbank 2021; Schierhorn et al. 2021), as well as drivers of natural hazards such as floods (Jiang et al. 2022) and wildfires (Richards et al. 2023; Bakke et al. 2023; Oliveira et al. 2012). Recently, researchers have called attention to contrasting, implausible, or ambiguous results (Lischeid et al. 2022; Mamalakis et al. 2023; Liu et al. 2022; Schmidt et al. 2020), suggesting that a deeper understanding of these methods is required. Known issues include disagreement between interpretation methods and the difficulty of satisfying the assumptions required for validity of results on spatiotemporal data. This is particularly relevant for agricultural studies, as due to the lack of long agricultural yield time series even in relatively data-rich regions, the use of data from multiple sites or regions is often necessary.

XAI is a rapidly advancing field, with scientific articles on the topic tripling in the last decade (Graziani et al. 2022), largely driven by research using established “benchmark” datasets that may have little similarity to climate or agricultural data. For researchers in Earth and climate sciences, a common challenge is the presence of autocorrelation in multivariate spatiotemporal data, while XAI methods often assume that data used are independent and identically distributed. A fundamental problem is robust model performance evaluation. It has been shown that the use of random cross validation to evaluate model skill on spatial, temporal, or spatiotemporal data, where the assumption of identically and independently distributed data is violated, can lead to overestimation of model performance (Meyer et al. 2019; Meyer and Pebesma 2021, 2022; Meyer et al. 2018; Vorndran et al. 2022; Kattenborn et al. 2022; Ploton et al. 2020; Hosseini et al. 2020; Roberts et al. 2017; Beigaitė et al. 2022). There is currently no established “best practice” for model evaluation on spatiotemporal climate data. Recent studies have made use of a range of methods, including a single defined test set, random *k*-fold, leave-one-year-out, or spatial cross validation, but methods that consider both the temporal and spatial dependencies of the dataset are rare (Richards et al. 2023; Roberts et al. 2017). Furthermore, the impact of the choice of evaluation strategy on the results of XAI methods is not yet known.

Using a methodological approach typical of studies aiming to identify climatic drivers of crop yield variability, we examine the impact of model evaluation strategy on the measured performance and interpretation, via permutation feature importances, of similar machine learning models. We further examine the impact of the evaluation strategy used for hyperparameter tuning and feature selection on the ability of the resulting models to extrapolate to years and regions that were not available during training and testing. By using simulated data from a process-based crop model, we are able to compare the interpretations derived from the machine learning models with the known mechanisms of the crop model and therefore comment on the plausibility of the “explanations.”

2. Data

The global climate dataset used consists of daily values of precipitation, minimum, maximum and average temperature and shortwave radiation, based on the NCEP–NCAR reanalysis product (Sheffield et al. 2006). The data are obtained from the Global Gridded Crop Model Intercomparison (GGCMI) phase 1 input datasets (Elliott et al. 2015) and cover the period 1948–2008 at $0.5° \times 0.5°$ resolution.

These data are used to drive the global gridded crop model ensemble of GGCMI phase 1 (Müller et al. 2017, 2019) and the Intersectoral Impact Model Intercomparison Project (ISIMIP) phase 2a initiative. Crop-planting date and growing-season length are prescribed at gridcell level, based on observational data of monthly irrigated and rainfed crop areas around the year 2000 (MIRCA2000; Portmann et al. 2010), as described in the modeling protocol (Elliott et al. 2015). Land use is not considered, but this analysis is limited to current cropping areas (defined according to MIRCA2000). We here use simulated maize yield data (tonnes dry matter ha^{-1}) for rain-fed systems from the Lund–Potsdam–Jena managed Land (LPJmL) model. In LPJmL, yield data are computed based on canopy-level photosynthesis, autotrophic respiration, limitations in water supply, and a set of allocation rules, driven by daily weather conditions and computed soil dynamics (Bondeau et al. 2007; Fader et al. 2010; Schaphoff et al. 2013; Waha et al. 2012). The version of LPJmL used for GGCMI phase 1 does not account for nitrogen stress, so fertilizer inputs as specified by the modeling protocol (Elliott et al. 2015) are ignored in the LPJmL simulations. As a robustness check, simulations from a global gridded modification of the Decision Support System for Agrotechnology Transfer (pDSSAT) crop model (Elliott et al. 2014; Jones et al. 2003), which did account for nitrogen fertilizer inputs, are also used.

The simulated maize yields are detrended and transformed into yield anomalies by fitting an order-3 polynomial and subtracting this from the yields at each grid point. An overview of the data-processing, model-training, and analysis steps is shown in Fig. 1.

For our climate predictors, we use monthly average precipitation p_i, temperature t_i, and shortwave radiation r_i for

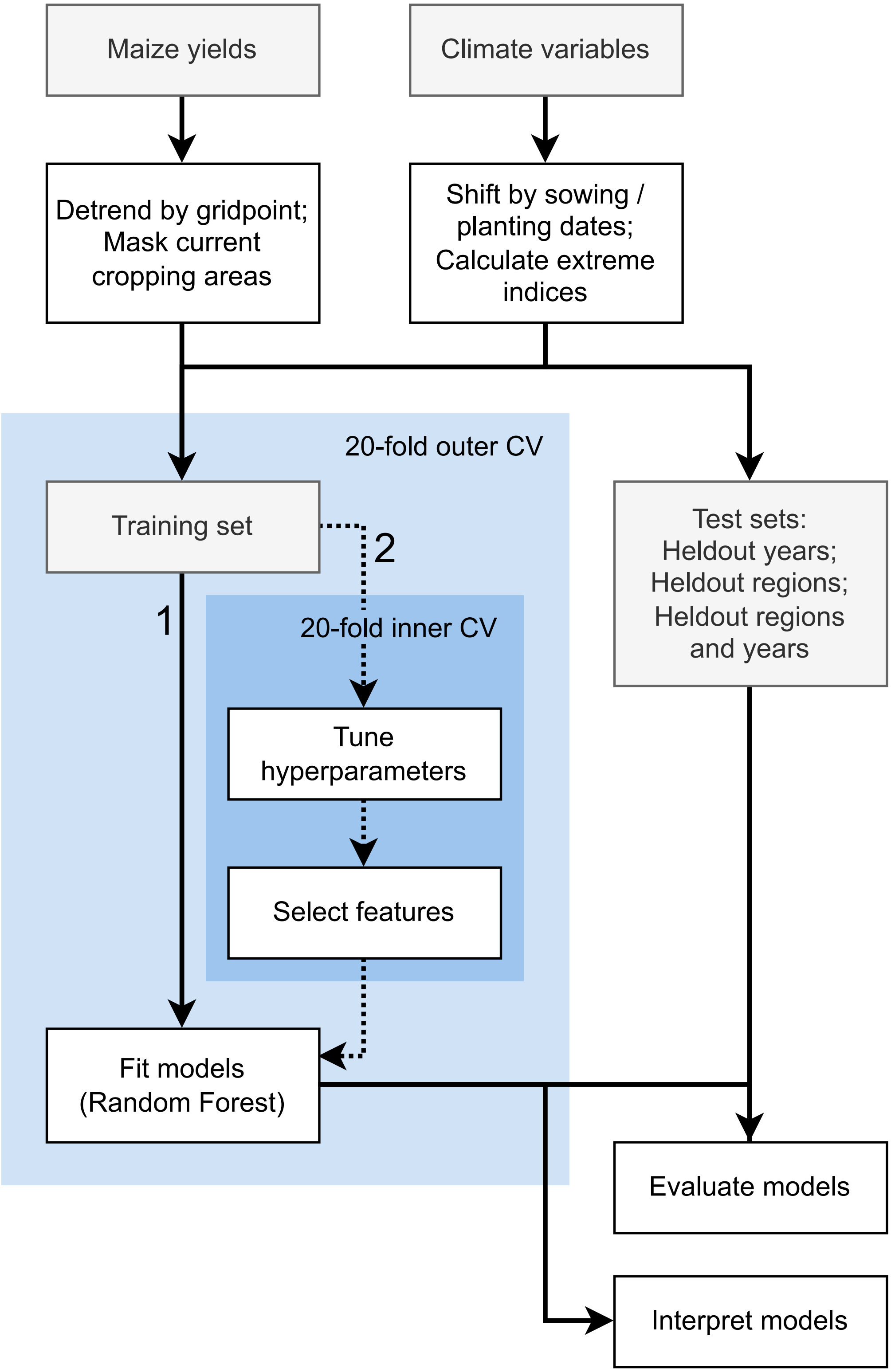
Maize yields
Climate variables
Detrend by gridpoint; Mask current cropping areas
Shift by sowing / planting dates; Calculate extreme indices
20-fold outer CV
Training set
2
1
20-fold inner CV
Tune hyperparameters
Select features
Fit models (Random Forest)
Test sets: Heldout years; Heldout regions; Heldout regions and years
Evaluate models
Interpret models

FIG. 1. Overview of the experimental workflow. In the first experiment (label 1), which aims to analyze the impact of the CV method on model evaluation and interpretation, no hyperparameter tuning or feature selection is conducted, and therefore the resulting models from each training fold are comparable. In the second experiment (label 2), which aims to find the impact of the CV method used on the ability of models to extrapolate to held-out years and/or regions, nested (inner) CV was used to tune hyperparameters and select features inside each outer training fold.

3 months prior to the sowing date and the subsequent 7 months, thereby fully encompassing the growing season in all grid cells (with subscript *i* used to represent the number of months before or after sowing). We additionally include a number of extreme indicators (Vogel et al. 2019) that are calculated over the period from sowing to maturity date. These indicators consist of the number of warm days (WD, defined as days above the 90th percentile for that grid cell); the number of cool days (CD, days below the 10th percentile for that grid cell); the number of frost days (FD, days below freezing); the maximum and minimum temperature experienced ($t_{\min}$ and $t_{\max}$, respectively); the maximum 5-day total precipitation (Rx5day); and the average diurnal temperature range (DTR). The resulting dataset comprises 37 predictive features.

We withhold the last 6 years of data, along with two spatial regions (selected from different continents and hemispheres; see Fig. S1 in the online supplemental material), in order to later evaluate model performance on years and locations to which it was not exposed during training. The spatial regions were selected arbitrarily. These held-out data are separated into three sets, henceforth referred to as the "held-out years," "held-out regions," and "held-out years and regions." The remaining 54 years and 30 034 grid points, which in total make up 1 546 580 datapoints, become the "training set" used for model fitting.

3. Method

We use a method similar to that used in recent studies attempting to identify drivers of agricultural yields with machine learning, and vary only the model evaluation strategy. While studies have employed a large variety of machine learning algorithms, random forests are frequently used for agricultural studies (Vishwakarma et al. 2022). In comparison with more complex architectures, random forests require little to no data preprocessing and are often able to achieve excellent performance on tabular data. We use the implementation from the sklearnex package for the Python programming language, an adaptation of scikit-learn (Pedregosa et al. 2011).

a. *Model evaluation*

Cross validation (CV) is widely used for robust machine learning model evaluation. To evaluate model performance, the training set is split into sections according to the chosen CV method. Random forest models are trained on the data in all but one of those sections (the "training fold") and tested on the single remaining section (the "test fold"). This process is repeated such that each section is used as a test fold once. The final model performance evaluation consists of the average of the resulting scores weighted by the number of datapoints in the respective test folds.

A larger number of CV folds allows more training data to be used for model fitting, but is more computationally intensive. At the most extreme, leave-one-out CV (LOOCV) is sometimes used, where the number of folds is equal to the number of datapoints. However, a more moderate number of folds (10–20) has been found to be better for model selection on real-world data (Kohavi 1995). For this study, we opt to use 20 CV folds.

In this study, we analyze the impact of the use of six different CV strategies. In each strategy, the training set datapoints (across all years and grid cells) are split into 20 folds, either at random or according to their spatial, temporal, or climatic characteristics. The six CV strategies studied are defined as follows. (i) random 20-fold CV, where datapoints in each fold are selected randomly from all years and grid cells of the training set; and (ii) temporal CV, where the training set is split by year into 20 consecutive folds. We define two spatial CV methods: (iii) latitude CV, where data are split by latitude into folds containing an equal number of datapoints, and (iv) spatially clustered CV, where datapoints are grouped into 20 clusters based on their latitude and longitude values using the *k*-means clustering algorithm. Similarly, for (v) spatiotemporally clustered CV, datapoints are clustered according to their latitude, longitude, and year. Finally, to consider the climatic variation in the dataset, we define (vi) "feature-clusters CV," where datapoints are clustered with the bisecting *k*-means algorithm on the 37 climate features previously described. Features were not standardized before clustering. Although no temporal or spatial information is explicitly given to the clustering algorithm, the resulting clusters (unsurprisingly) exhibit spatiotemporal patterns.

As the metric of model performance, we select the coefficient of determination (R^2) due to its frequent use in similar studies. R^2 represents the proportion of the variance explained by the variables of the model. The best possible score is 1.0, and a model that predicts the mean of the target variable would have an R^2 of 0.0; R^2 can also be negative. Our results are robust to the use of explained variance or root-mean-square error (RMSE) as performance metric.

b. *Model interpretation*

In this study, we focus on permutation feature importance, which is one of the most widely used model interpretation strategies. Permutation feature importances are model agnostic and do not require retraining (in other words, they are a post hoc XAI method). The results are highly human comprehensible.

The model performance is first evaluated on a test set (or CV test fold). To identify the importance of a feature, the values of that feature are randomly shuffled (permuted) over the test set, thereby destroying any relationship between the feature and the target variable. The model performance is then evaluated on the perturbed test set, and the difference in score before and after permutation is the permutation feature importance. This process is often repeated a number of times for robustness. When using CV, this process is additionally repeated for each test fold.

c. *Experiment 1: The effect of CV on model evaluation and interpretation*

We first examine the impact of CV strategy on the outcome of model evaluation and interpretation, while keeping models identical in terms of hyperparameters (Table 1) and predictive features. For each of the 20 CV folds, we train one

TABLE 1. Random forest hyperparameter values used. In the first experiment, random forest hyperparameters are held at the given values. In the second experiment, hyperparameters are selected from the ranges defined during two tuning stages conducted before and after feature selection.

Hyperparameter	1	2.1	2.2
No. of trees	400	10–500	10–500
Max tree depth	100	2–120	2–120
Min samples to split node	1	1–30	1–30
Max features considered	0.5	1.0	0.2–1.0

random forest regression model on all training fold data (which contain multiple years and grid cells). The model is trained to predict maize yield anomaly using the 37 climate features described above as predictors. We then measure the model performance and calculate permutation feature importances (repeating 10 times and averaging for robustness) on the test fold data. This methodology is repeated for each CV strategy tested.

To assess the accuracy of the model performance scores returned, model skill is then evaluated on the held-out regions, held-out years, and held-out years and regions after retraining the model on the entire training set.

d. *Experiment 2: The effect of CV on model skill on held-out years and/or regions*

We next investigate the impact of CV strategy on model performance, when used during all stages of a machine learning pipeline. In each training fold, hyperparameter tuning and feature selection are conducted using a second 20-fold CV process (nested CV). The resulting 20 models can therefore vary in model structure and features used. These models are then evaluated on the respective test folds and on the held-out regions and/or years.

Hyperparameter tuning is conducted using the Optuna package (Akiba et al. 2019), with hyperparameters sampled from distributions described in Table 1. After a first tuning stage, a minimum of five and maximum of 20 features are selected using sequential forward floating selection (SFFS) (Pudil et al. 1994). In this process, the single predictive feature is selected that provides the best performance (using inner 20-fold CV). Features are then iteratively added according to which addition results in the best model performance. After each step, model performance after removal of each feature is tested and, if resulting in model improvement, executed. After feature selection is complete, a second tuning stage is conducted, and the models are evaluated and interpreted as in the previous experiment.

4. Results

a. *Cross-validation impact on model evaluation*

Performance scores obtained on the training set for random forest models trained to predict maize yield anomalies using all 37 climate features, with no hyperparameter tuning, vary widely by the CV strategy used (Fig. 2). Using random 20-fold CV returns an estimated R^2 of 0.82, with very little variation between folds. Using temporal, spatial, spatiotemporal, or feature-clusters CV returns lower estimations of model performance (albeit with higher variation between folds).

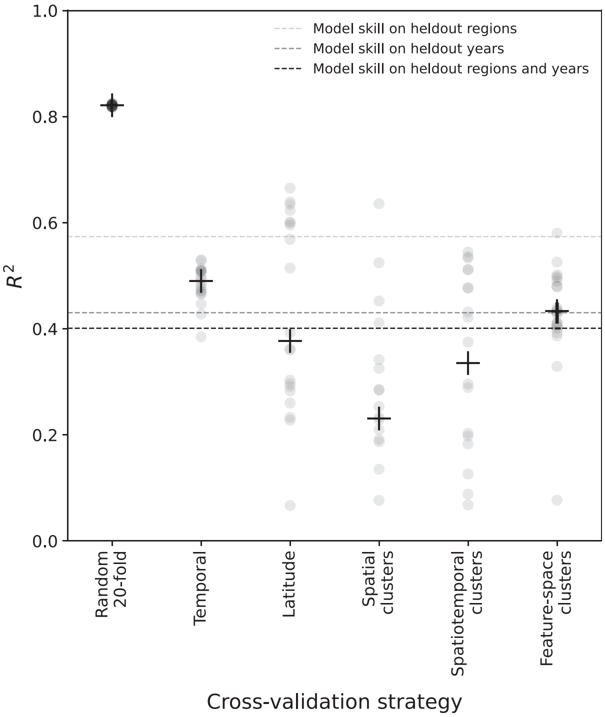

FIG. 2. Model skill as evaluated using various CV strategies on the training set. Each point represents the performance as measured using one CV test fold, and crosses represent the median score. Horizontal dashed lines denote model skill (after retraining on the entire training set) on held-out years and regions.

After retraining on the entire training set, the random forest model achieves R^2 scores of 0.43 on held-out years, 0.57 on held-out regions, and 0.40 on held-out years and held-out regions. The difference between the score achieved on the held-out data and the scores calculated using CV on the training set is most extreme when using random 20-fold CV, for which the R^2 on held-out years and regions is less than half of the CV-estimated score.

Similar results are obtained for the pDSSAT crop model output data, using identical methods and model specifications (Fig. S2 in the online supplemental material), although model performance is poorer overall. On held-out years and regions, R^2 of 0.25 is achieved; using 20-fold CV on the training set to estimate model skill returns a much higher R^2 of 0.62.

b. *Cross-validation impact on model interpretation*

Using all CV strategies, precipitation in the second and third month after sowing (p_{+2} and p_{+3}) are identified as important features for the prediction of yield anomalies, while the precipitation in months outside of this period shows comparatively lower importance (Fig. 3c). Similarly, r_{+3} is identified as an important feature using all CV strategies (Fig. 3b),

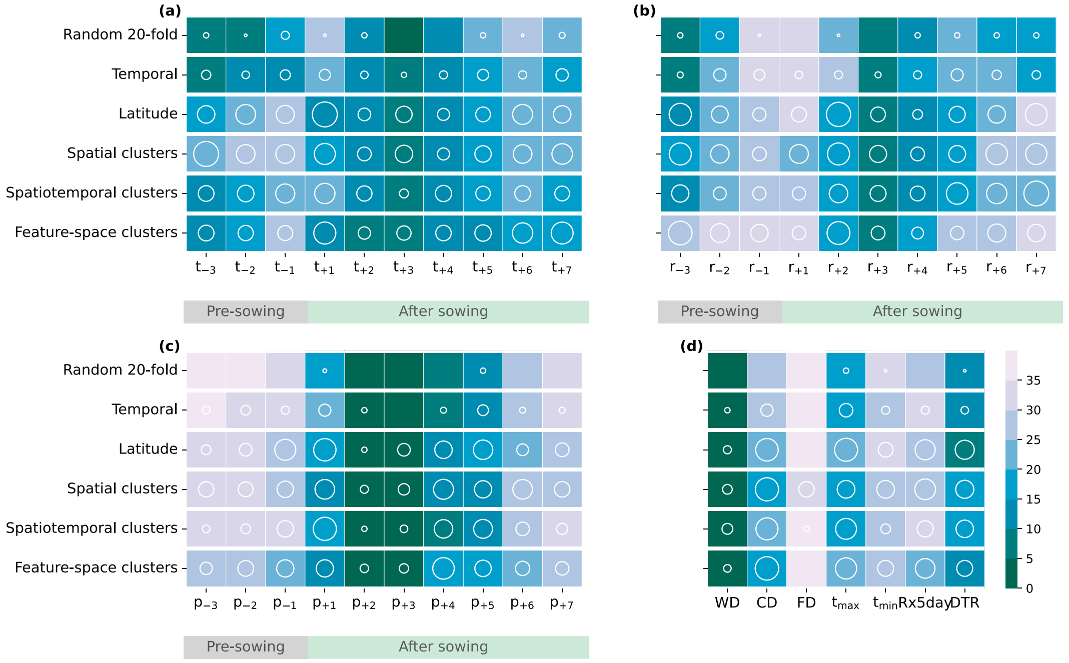

FIG. 3. Ranked permutation feature importances of the random forest models, calculated using different CV strategies, for the following 37 climatic features: average monthly (a) temperature, (b) radiation, and (c) precipitation from 3 months before until 7 months after sowing, as well as (d) the number of warm days during the growing season (WD), cool days (CD), frost days (FD), the maximum and minimum daily temperature (t_{min} and t_{max}), the maximum 5-day total precipitation (Rx5day), and the average diurnal temperature range (DTR). The t_i, r_i, and p_i represent the average monthly temperature, radiation, and precipitation during the ith month after sowing, respectively. Shading shows the median ranked importance across 20 folds, and size of circles shows a qualitative measure of the range in feature importances across test folds, where larger circles indicate higher variability between the folds.

as well as WD (Fig. 3d). Some features return comparatively low feature-importance scores across all strategies, such as FD and the precipitation, temperature, and radiation later than four months after sowing.

Variation between the permutation feature importances for each fold is lowest when using random 20-fold CV (size of circles in Fig. 3). Use of temporal CV also results in relatively little variation between the folds, whereas variation was highest when using spatially clustered CV.

In general, the difference between feature importances is highest between random 20-fold CV and feature-clusters CV. This is evidenced by a Spearman correlation coefficient of 0.71 between the median feature importances in each CV F4 test fold (Fig. 4). In contrast, the correlation between feature importances calculated using random 20-fold and temporal CV is 0.99.

The permutation feature importances are strongly affected by the CV method, particularly for r_{-3} and t_{-3}. These features are identified as important when using random 20-fold or temporal CV (in fact, r_{-3} is found to be the most important month of radiation). However, these features are estimated to have low importance when using other CV strategies (with t_{+3} and r_{+3} instead identified as the most important months). Although variation between folds is higher for these methods, these features have low importance in all CV folds.

To further explore the importance of the climate conditions in the third month prior to sowing, we remove these features from the dataset and repeat the model evaluation and interpretation procedure with the 34 remaining features (Fig. S3 in

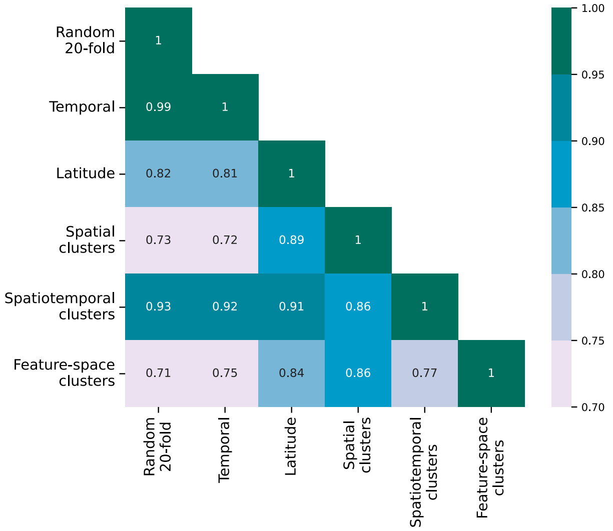

FIG. 4. Spearman correlation coefficient calculated between the median permutation feature importances in each test fold, using each CV strategy.

the online supplemental material). Given that these features have high importance when using random 20-fold or temporal CV, a decrease in model performance could be expected; however, model performance is found to be only slightly lower on the training set and, in fact, slightly higher on the held-out years and regions.

While model hyperparameters are kept constant in this experiment, the use of different CV strategies implies slight variation of the training folds used for each model and therefore the data used for training, which could potentially result in different model structures. As 20 folds are used, the difference in training folds is around 5% of the datapoints. Comparing the models directly is infeasible, but as an indication of model similarity, we calculate the internal feature importances, which do not depend on a chosen test set, and find them to be near identical (Fig. S4 in the online supplemental material).

c. *Impact of CV during hyperparameter tuning and feature selection*

The CV strategy chosen is found to have an impact on both the values of the hyperparameters chosen when tuning and the features selected using SFFS (Figs. S5 and S6 in the online supplemental material). Interestingly, even the very first feature selected varies depending on the CV strategy. Using random 20-fold or temporal CV leads to p_{+3} being selected first for all folds, while r_{+2} is selected when using latitude or feature-clusters CV. Apart from spatially clustered CV, use of all strategies results in the maximum possible number of features (20) being selected. The final features selected varies the least between folds when using random 20-fold or temporal CV, with 17 and 16 of the 20 features selected identical for all folds.

d. *Impact on model performance on held-out years and/or regions*

The resulting models are then evaluated on the held-out years and regions (Fig. 5). The use of feature-cluster or temporal CV results in better model skill on held-out years and regions (median $R^2 = 0.42$) than the use of random 20-fold CV (median $R^2 = 0.37$), and the use of spatial or spatiotemporal CV leads to worse model performance. The variation in model performance scores between folds is lowest for random 20-fold CV.

Using spatially clustered CV leads to much poorer model skill on the held-out sets than the other strategies tested. This may be in part due to the upper bound (20) set on the number of features selected. Inspection of the trend of the performance of the models against the number of features used suggests that setting a higher bound may have offset this difference somewhat.

Results are similar when evaluating model performance on held-out years or held-out regions (Fig. S7 in the online supplemental material). Model performance is worse after tuning and feature selection, with temporal CV resulting in the best model performance on held-out years and feature-clusters CV resulting in the best model performance on held-out regions. The general decrease in model performance after these steps could arise from our choice to set the maximum number of features to 20, which may limit the achievable model performance.

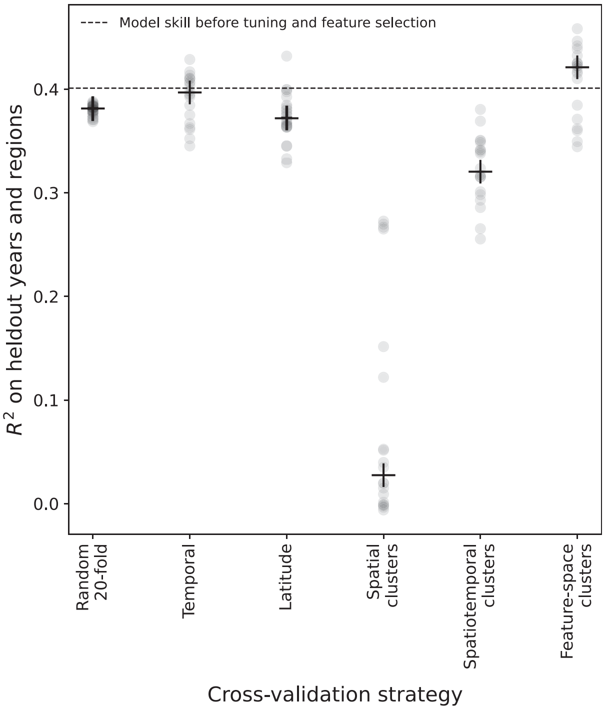

FIG. 5. Model skill on held-out years and regions, after hyperparameter tuning and feature selection is conducted using six different CV strategies. Each point represents the performance of one resulting model that was trained on one CV training fold, and crosses denote the median performance. The dashed line marks the model performance from the previous experiments, where no tuning or feature selection was done and the model was retrained on the entire training set.

Permutation feature importances of the final models using the different CV strategies (Fig. S8 in the online supplemental material) show similar characteristics to those obtained in the previous experiment. Using spatially clustered CV results in very different permutation feature-importance results. However, as the performance of these models is very low on both the test folds of the training set and the held-out regions and/or years (Fig. 5, along with Fig. S7 in the online supplemental material), these results cannot be meaningfully interpreted.

5. Discussion

a. *Cross-validation strategy used has an impact on model evaluation*

The use of interpretable or explainable machine learning tools for the identification of drivers of agricultural impacts has become increasingly widespread (Goulart et al. 2021; Peichl et al. 2021; Webber et al. 2020; Wolanin et al. 2020; Mateo-Sanchis et al. 2021; Martínez-Ferrer et al. 2020; Newman and Furbank 2021; Schierhorn et al. 2021). Given the complex relationships between climate, crop phenology, and yields (Lesk et al. 2022; Schauberger et al. 2016) and the growing availability of relevant data, this trend seems likely to continue. Researchers have identified pitfalls in the use of XAI methods for research in geosciences (Mamalakis et al. 2022, 2023), and, in parallel, studies have called attention to issues arising from the use of random k-fold CV for model evaluation on spatial or temporal ecology or climate data (Meyer et al. 2019; Meyer and Pebesma 2022; Vorndran et al. 2022; Kattenborn et al. 2022). Our results provide further evidence that random k-fold CV may overestimate model skill in comparison with performance on held-out spatial regions and/or years (Fig. 2). In our study, the estimated R^2 score using random k-fold CV on the training set is over double that measured on held-out years and regions.

An accurate measure of skill is necessary when training predictive models in order to assess if a sufficient performance threshold has been reached. It is also essential when trained models are instead used for analysis via XAI methods. The model must successfully capture the underlying data-generation processes in order for those processes to then be identified using XAI (Molnar et al. 2022; Jiang et al. 2022). Good predictive performance is therefore needed, although it is alone not proof that the model has captured true causal drivers. Thus, this finding indicates a need for the definition of best practices for the methodology of similar studies, which would depend on both the intended use of the model (prediction or interpretation) and the characteristics of the dataset used.

b. *Cross-validation strategy used has an impact on model interpretation*

In studies making use of XAI methods to understand physical drivers of impacts, models are rarely intended for use on data from future years or additional spatial regions. The appropriate choice of cross-validation method may therefore be seen as less relevant, provided that the model achieves good performance on the studied distribution. However, our finding that the cross-validation method used has an impact on the calculated permutation feature importances indicates that the choice of model evaluation strategy is, in fact, highly relevant for such studies.

We find that the largest disparity in resulting ranked permutation feature importance (in terms of the Spearman correlation coefficient) occurs between random 20-fold CV and feature-clusters CV (Fig. 4). A direct comparison of the ranked feature importances (Fig. 6) shows that the feature with the largest discrepancy is r_{-3}. Using random 20-fold CV, r_{-3} is identified as the most important month of radiation and fifth-most-important feature overall. However, it falls to the bottom third of predictive features when using feature-clusters CV, with r_{+3} instead identified as the highest-importance month of radiation and fifth-most-important feature overall. Furthermore, the other months of temperature and radiation presowing are substantially less important using feature-cluster CV than with random 20-fold CV, while the presowing precipitation months are of higher importance using feature-cluster CV (albeit in the bottom half of all features).

Crop growth and final yield are mainly determined by weather conditions during the growing period in a diversity of processes (Schauberger et al. 2016). Soils can, to some extent, represent weather conditions from before the growing season by storing water and heat. Nutrients stored in soils can also be affected by pregrowing weather (Li et al. 2022), but nitrogen dynamics are not considered by the LPJmL simulation data used here. Preseason radiation and temperatures can affect soil water and temperature at the beginning of the growing season. However, in general, conditions closer to the growing season should be more important than those farther away, and conditions during the growing season are substantially more important than conditions outside the growing season, for example, by directly affecting photosynthesis and autotrophic respiration. This suggests that the feature importances calculated using feature clusters are a more plausible reflection of the underlying data-generating model.

Further evidence for this is provided by the fact that when r_{-3}, t_{-3}, and p_{-3} are removed from the dataset, model performance scores on held-out years and regions increase slightly, while scores on the training set decrease (Fig. S3 in the online supplemental material).

We observe differing amounts of variation in permutation feature importances between CV folds for each CV strategy. This variation is often interpreted as a measure of uncertainty, which would imply that the interpretations obtained using random 20-fold CV carry the least uncertainty. However, the validity of reducing highly complex, nonlinear models to human-intelligible "explanations" via XAI methods has been questioned (Molnar et al. 2022), and we hypothesize that the variation in importances between folds may express interacting relationships captured by the model.

For example, we calculate the average value of WD for all datapoints in each test fold when using feature-clusters CV and compare this with the calculated permutation feature importance of WD in that fold (Fig. 7). The importance of WD steadily increases as the datapoints in the folds become warmer until a maximum of 40 is reached. At that point, the importance varies (possibly depending on other climate characteristics of the test folds). This type of analysis is possible when calculating permutation feature importance on diverse folds in feature space, but not when using random k-fold CV, as the datapoints in each test fold will be similarly distributed.

Some climate features are identified as having relatively high (or low) permutation feature importance using all CV strategies tested, which suggests that repeating the calculation of feature importance using multiple methods could provide more confidence in the interpretations.

c. *Cross-validation strategy used has an impact on model performance on held-out years and regions*

Our results show that the choice of cross-validation strategy not only is relevant for studies using XAI, but also has implications for the predictive skill of the model. The hyperparameter values chosen and features selected are found to vary depending on the inner 20-fold CV strategy used (Figs. S5

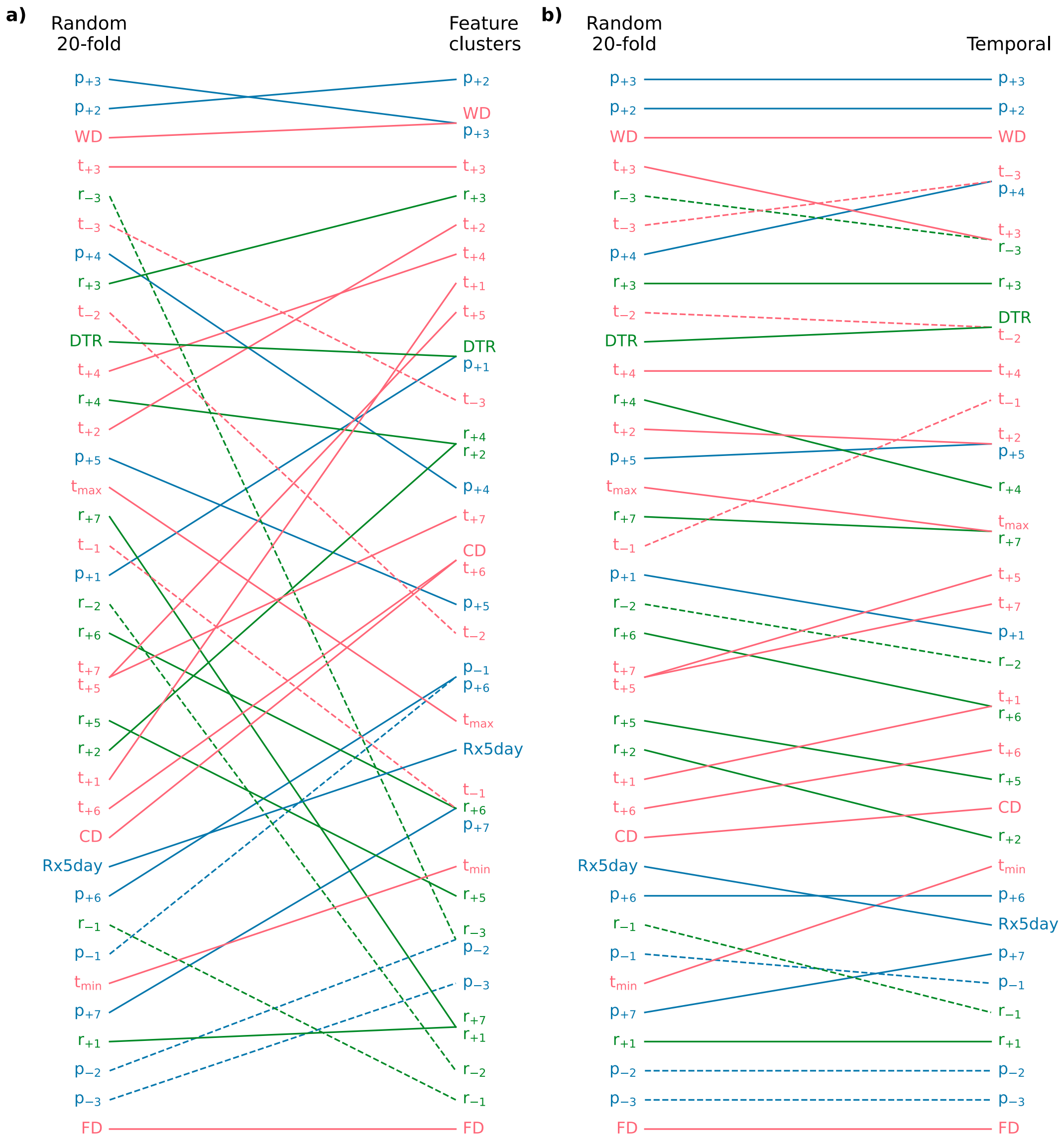

FIG. 6. A direct comparison of the ranked feature-importance scores using random 20-fold CV vs (a) features-clusters CV or (b) temporal CV. High-importance features are at the top, and low-importance features are at the bottom. The dashed lines denote features representing climatic conditions before sowing, and colors represent the related climate variable (precipitation, temperature, or radiation).

and S6 in the online supplemental material), ultimately leading to models that vary in skill when evaluated on held-out years and regions (Fig. 5). Restricting the maximum number of features to 20 leads to decreased model skill for all evaluation strategies except feature-clusters CV. Additionally, the use of temporal CV returns models with better skill than those created using random 20-fold CV. Tests on held-out years and held-out regions (Fig. S7 in the online supplemental material) show similar results, with the use of evaluation strategies such as temporal CV leading to better model skill than random *k*-fold.

The improved model skill when using feature-clusters CV for feature selection, despite restricting the number of predictors to 20, suggests that the features chosen better reflect the structure of the underlying process-based model.

d. Feature-clusters cross validation

Permutation feature importances, in common with all current XAI methods, have numerous pitfalls (Molnar et al. 2022). Most relevant to studies using climate-related data is their sensitivity to correlations between predictive features (Hooker et al. 2021). One method sometimes used to remediate this issue is to group correlated features and then permute grouped features collectively. This is challenging, however, for high-dimensional datasets where the majority of or all features are correlated. Furthermore, the bivariate correlations

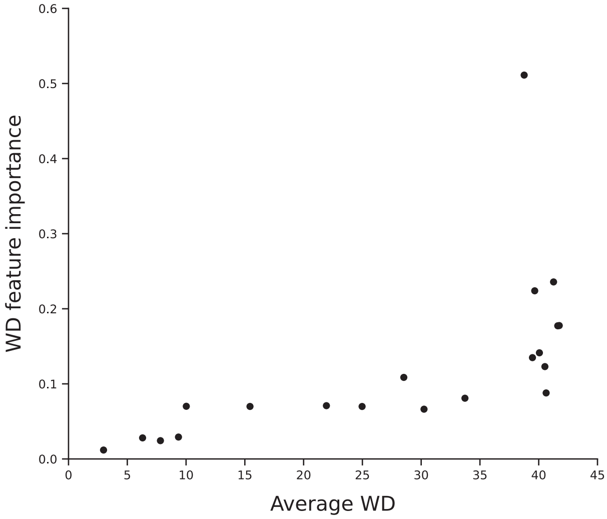

FIG. 7. For each of the 20 CV folds here (split by clusters in feature space), dots denote the permutation feature-importance score of the number of WD during the growing season against the mean WD in that CV fold.

between features may not capture all relationships in the multivariate joint distribution. A second strategy is "leave one feature out" feature importance (LOFO), where the model is retrained with one feature omitted, and the difference in model performance is interpreted as the importance. This requires retraining the model for each feature (which can be computationally expensive) and therefore compares two different prediction models rather than evaluating a single model. "Conditional permutation feature importances" have also been proposed, where the shuffling of each feature is done with respect to the joint distribution (Fisher et al. 2019; Strobl et al. 2008), but methods are either model specific or infeasible for high-dimensional datasets. In practice, these methods are rarely used in applied research, in contrast to permutation feature importances.

Our data exhibit correlations between most features. Under these conditions, permuting generates artificial datapoints that are far from the joint distribution (Hooker et al. 2021). For example, in our dataset, this process could result in models being evaluated on artificial datapoints where the number of frost days was above zero but the minimum temperature reached during the growing season was considerably above freezing levels. Such datapoints are physically implausible. This issue can be tackled by the use of so-called knockoffs, artificial datapoints that preserve the multivariate correlation structure of the dataset (Barber and Candès 2015). Generation of knockoffs is difficult in high-dimensional scenarios, but generative deep learning models may be used (Jordon et al. 2019; Romano et al. 2020).

Using feature-clusters CV to calculate permutation feature importance means that datapoints are permuted across the values in one cluster only. We find that this restricts the generation of such implausible datapoints, in comparison with the use of random k-fold CV (Fig. 8). In fact, we find that the use of feature-clusters CV preserves a large number of the correlations between features in the artificial permuted datapoints (Fig. S9 in the online supplemental material), and we are thereby implicitly creating knockoffs.

However, this does not appear to be the cause for the difference in feature importances between random 20-fold CV and feature-clusters CV. To investigate this, we create knockoff datapoints using feature-clusters CV and then calculate feature importance by substituting each feature's data with the knockoff data and measuring the difference in model skill using random 20-fold CV. If the preserved correlation structure is driving the disparity in feature importances, we would expect the importances to differ from those returned from random 20-fold CV previously. However, the feature importances are similar to those originally obtained (Fig. S10 in the online supplemental material).

A second hypothesis for the divergence in interpretations is that when using feature-clusters CV, the model is forced to extrapolate in feature space when predicting the datapoints in each test fold. Features that allow the model to overfit to the distribution of the training fold (due to spurious correlations) would not improve model skill and therefore have low importance. High-importance features are those that generalize and so are more likely to reflect the underlying data-generating process. This may explain the greater plausibility of the interpretations when using feature-clusters CV. In addition, this would account for the comparatively minor differences between the ranked feature importances obtained when using random k-fold and temporal CV; the interannual variation in climate is relatively small, and therefore, the model is not extrapolating when making predictions on the test folds.

This hypothesis suggests that the best choice of cross-validation strategy for a study using XAI will depend on the intended use of the interpretations. The fact that the feature importances returned using random k-fold CV were not changed by using the knockoff datapoints suggests that those interpretations may accurately reflect the machine learning model. If the purpose of the interpretations is to better understand the model in order to identify bias or for model diagnostics, this may be the optimal strategy. For studies attempting to better understand the physical data-generating process, the use of feature-clusters CV may omit features that do not generalize to held-out environments and are therefore less likely to be true causal drivers.

An alternative explanation for the disparity in interpretations is that the features found to have high importance using random 20-fold CV but low importance using feature-clusters CV (e.g., r_{-3} and t_{-3}) may be good predictors of yield anomalies on a global level, but bad predictors "locally," in comparison with other features. However, these features have comparatively low correlation with the target variable (maize anomalies), and therefore, this predictive skill must depend on interaction with other variables.

This study does not examine the impact of cross validation on other XAI methods such as Shapley values. Furthermore, our study is restricted to random forest models. Other studies have found similar issues with permutation importances for neural networks (Hooker et al. 2021), but extending this study

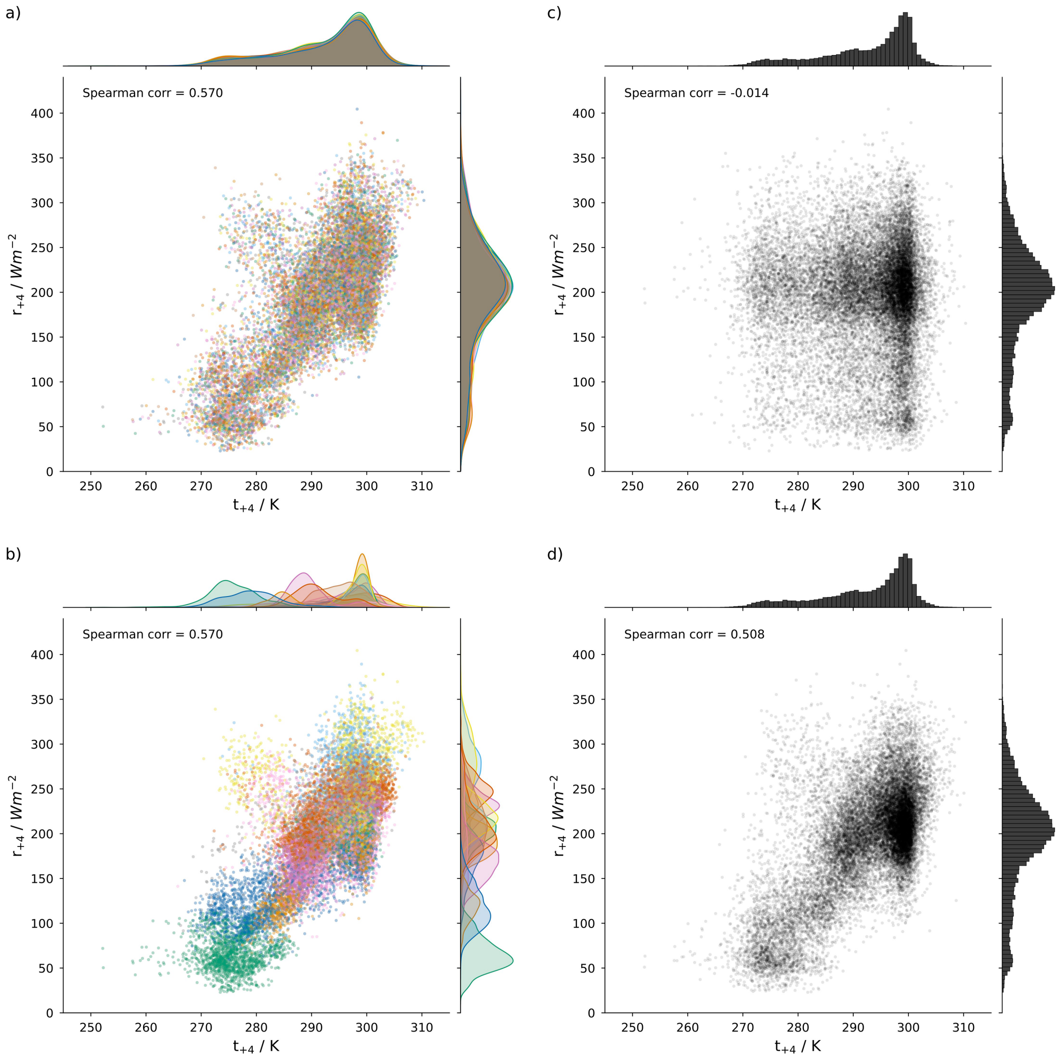

FIG. 8. An illustration of the artificial datapoints used when calculating permutation feature importance using either random k-fold CV or feature-clusters CV, for two highly correlated features (t_{+4} and r_{+4}). Shown is a sample of 20 000 datapoints from the training set, colored by CV fold used via (a) random 20-fold CV and (b) feature-clusters CV. Also shown are (c),(d) the same datapoints after permuting one feature in each fold using the respective CV strategy (note that the outcome is identical regardless of which of the two features is permuted). The process using random 20-fold CV is shown in (a) and (c), and the use of feature-clusters CV is shown in (b) and (d). Using feature-clusters CV to generate permuted datapoints results in a distribution far closer to that of the original datapoints, which is also seen in the resulting Spearman correlation between the two features.

to other algorithms would increase confidence in the general applicability of our findings.

The use of simulated data from a process-based crop model, driven by the climate forcing data and without the use of fertilization, irrigation, or varying management practices, ensures that the underlying data-generating process is identical across regions and years. Given this idealized setting, we expect that the result that cross validation has an impact on the evaluation and interpretation of machine learning models can be assumed to be valid for other spatiotemporal datasets, including observational datasets. However, the precise nature of that impact will depend on the characteristics of the dataset and data-generating process. Future research could (i) make use of toy models where the data-generating processes is known in order to better understand the causes of this effect and (ii) apply similar methodology to simulated datasets with other types of correlation structures and/or observational datasets where the underlying process is well understood.

6. Conclusions

In this study, we directly compare the impact of six model evaluation strategies, including random k-fold and temporal and spatial CV methods, as well as feature-clusters CV in which datapoints are clustered in the space of the predictive features. We demonstrate that model evaluation using random *k*-fold CV may severely overestimate model skill on spatiotemporally correlated climate data, in agreement with existing research.

Our results show that the chosen CV strategy affects the permutation feature importances. By using simulated maize yield data from a process-based crop model as the target variable, we are able to comment on the plausibility of the explanations provided. We find that the use of random *k*-fold CV returns the least plausible feature importances, and feature-clusters CV the most. Importantly, although using temporal CV may give a more accurate estimation of model predictive skill on held-out years, the feature importances are very similar to those calculated using random *k*-fold CV (Spearman correlation coefficient = 0.99). This suggests that the use of temporal CV may be sufficient for estimation of model predictive power in some cases, but it is not the optimal choice when using XAI to study a physical process.

Last, we show that the model evaluation strategy used during hyperparameter tuning and feature selection has an impact on the skill of models on held-out years and spatial regions. Our results suggest that careful selection of CV strategy may improve the generalizability of the model. This is particularly relevant for agricultural studies, where data availability and quality vary across countries and regions.

Overall, our results demonstrate that the choice of cross-validation strategy has an impact on the outcome of model interpretation as well as model evaluation metrics, and therefore must be carefully chosen when conducting research using machine learning on spatiotemporal climate data. Our findings provide first steps toward establishing best practices for this choice in future studies.

Acknowledgments. Author Sweet, Anand, and Zscheischler acknowledge funding from the Helmholtz Initiative and Networking Fund (Young Investigator Group COMPOUNDX, Grant Agreement VH-NG-1537).

Data availability statement. All data used in this study are openly available. The simulated maize yields along with the planting and maturity dates from LPJmL can be accessed at Zenodo (https://zenodo.org/record/1403073#.ZBnkaOzMK3I) or at the ISIMIP repository, along with the atmospheric climate input data (https://doi.org/10.48364/ISIMIP.886955). Code used to generate the results from this study is available online (https://doi.org/10.5281/zenodo.7967133).

REFERENCES

Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama, 2019: Optuna: A next-generation hyperparameter optimization framework. *KDD'19: Proc. 25rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Anchorage, AK, Association for Computing Machinery, 2623–2631, https://doi.org/10.1145/3292500.3330701.

Bakke, S. J., N. Wanders, K. van der Wiel, and L. M. Tallaksen, 2023: A data-driven model for Fennoscandian wildfire danger. *Nat. Hazards Earth Syst. Sci.*, **23**, 65–89, https://doi.org/10.5194/nhess-23-65-2023.

Barber, R. F., and E. J. Candès, 2015: Controlling the false discovery rate via knockoffs. *Ann. Stat.*, **43**, 2055–2085, https://doi.org/10.1214/15-AOS1337.

Beigaitė, R., M. Mechenich, and I. Žliobaitė, 2022: Spatial cross-validation for globally distributed data. *Discovery Science*, P. Pascal and D. Ienco, Eds., Lecture Notes in Computer Science, Vol. 13601, Springer, 127–140, https://doi.org/10.1007/978-3-031-18840-4_10.

Ben-Ari, T., J. Boé, P. Ciais, R. Lecerf, M. Van der Velde, and D. Makowski, 2018: Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. *Nat. Commun.*, **9**, 1627, https://doi.org/10.1038/s41467-018-04087-x.

Bondeau, A., and Coauthors, 2007: Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Global Change Biol.*, **13**, 679–706, https://doi.org/10.1111/j.1365-2486.2006.01305.x.

Brás, T. A., J. Seixas, N. Carvalhais, and J. Jägermeyr, 2021: Severity of drought and heatwave crop losses tripled over the last five decades in Europe. *Environ. Res. Lett.*, **16**, 065012, https://doi.org/10.1088/1748-9326/abf004.

Crane-Droesch, A., 2018: Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environ. Res. Lett.*, **13**, 114003, https://doi.org/10.1088/1748-9326/aae159.

Elliott, J., D. Kelly, J. Chryssanthacopoulos, M. Glotter, K. Jhunjhnuwala, N. Best, M. Wilde, and I. Foster, 2014: The parallel System for Integrating Impact Models and Sectors (pSIMS). *Environ. Modell. Software*, **62**, 509–516, https://doi.org/10.1016/j.envsoft.2014.04.008.

——, and Coauthors, 2015: The Global Gridded Crop Model Intercomparison: Data and modeling protocols for phase 1 (v1.0). *Geosci. Model Dev.*, **8**, 261–277, https://doi.org/10.5194/gmd-8-261-2015.

Fader, M., S. Rost, C. Müller, A. Bondeau, and D. Gerten, 2010: Virtual water content of temperate cereals and maize: Present and potential future patterns. *J. Hydrol.*, **384**, 218–231, https://doi.org/10.1016/j.jhydrol.2009.12.011.

Fisher, A., C. Rudin, and F. Dominici, 2019: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, **20**, 1–81.

Goulart, H. M. D., K. van der Wiel, C. Folberth, J. Balkovič, and B. van den Hurk, 2021: Storylines of weather-induced crop failure events under climate change. *Earth Syst. Dyn.*, **12**, 1503–1527, https://doi.org/10.5194/esd-12-1503-2021.

Graziani, M., and Coauthors, 2022: A global taxonomy of interpretable AI: Unifying the terminology for the technical and social sciences. *Artif. Intell. Rev.*, **56**, 3437–3504, https://doi.org/10.1007/s10462-022-10256-8.

Hamed, R., A. F. Van Loon, J. Aerts, and D. Coumou, 2021: Impacts of compound hot–dry extremes on US soybean yields. *Earth Syst. Dyn.*, **12**, 1371–1391, https://doi.org/10.5194/esd-12-1371-2021.

Heinicke, S., K. Frieler, J. Jägermeyr, and M. Mengel, 2022: Global gridded crop models underestimate yield responses to droughts and heatwaves. *Environ. Res. Lett.*, **17**, 044026, https://doi.org/10.1088/1748-9326/ac592e.

Hooker, G., L. Mentch, and S. Zhou, 2021: Unrestricted permutation forces extrapolation: Variable importance requires at least one more model, or there is no free variable importance. *Stat. Comput.*, **31**, 82, https://doi.org/10.1007/s11222-021-10057-z.

Hosseini, M., M. Powell, J. Collins, C. Callahan-Flintoft, W. Jones, H. Bowman, and B. Wyble, 2020: I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data. *Neurosci. Biobehav. Rev.*, **119**, 456–467, https://doi.org/10.1016/j.neubiorev.2020.09.036.

Iizumi, T., and N. Ramankutty, 2016: Changes in yield variability of major crops for 1981–2010 explained by climate change. *Environ. Res. Lett.*, **11**, 034003, https://doi.org/10.1088/1748-9326/11/3/034003.

Jägermeyr, J., and Coauthors, 2021: Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. *Nat. Food*, **2**, 873–885, https://doi.org/10.1038/s43016-021-00400-y.

Jiang, S., E. Bevacqua, and J. Zscheischler, 2022: River flooding mechanisms and their changes in Europe revealed by explainable machine learning. *Hydrol. Earth Syst. Sci.*, **26**, 6339–6359, https://doi.org/10.5194/hess-26-6339-2022.

Jones, J. W., and Coauthors, 2003: The DSSAT cropping system model. *Eur. J. Agron.*, **18**, 235–265, https://doi.org/10.1016/S1161-0301(02)00107-7.

Jordon, J., J. Yoon, and M. van der Schaar, 2019: KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks. *Seventh Int. Conf. on Learning Representations*, New Orleans, LA, ICLR, 1–25, https://openreview.net/pdf?id=ByeZ5jC5YQ.

Kattenborn, T., F. Schiefer, J. Frey, H. Feilhauer, M. D. Mahecha, and C. F. Dormann, 2022: Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open J. Photogramm. Remote Sens.*, **5**, 100018, https://doi.org/10.1016/j.ophoto.2022.100018.

Kohavi, R., 1995: A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 2, Morgan Kaufmann Publishers Inc., 1137–1143, https://dl.acm.org/doi/10.5555/1643031.1643047.

Lafferty, D. C., R. L. Sriver, I. Haqiqi, T. W. Hertel, K. Keller, and R. E. Nicholas, 2021: Statistically bias-corrected and downscaled climate models underestimate the adverse effects of extreme heat on U.S. maize yields. *Commun. Earth Environ.*, **2**, 196, https://doi.org/10.1038/s43247-021-00266-9.

Laudien, R., B. Schauberger, D. Makowski, and C. Gornott, 2020: Robustly forecasting maize yields in Tanzania based on climatic predictors. *Sci. Rep.*, **10**, 19650, https://doi.org/10.1038/s41598-020-76315-8.

Lesk, C., W. Anderson, A. Rigden, O. Coast, J. Jägermeyr, S. McDermid, K. F. Davis, and M. Konar, 2022: Compound heat and moisture extreme impacts on global crop yields under climate change. *Nat. Rev. Earth Environ.*, **3**, 872–889, https://doi.org/10.1038/s43017-022-00368-8.

Li, Z., and Coauthors, 2022: Assessing the impacts of pre-growing-season weather conditions on soil nitrogen dynamics and corn productivity in the U.S. Midwest. *Field Crops Res.*, **284**, 108563, https://doi.org/10.1016/j.fcr.2022.108563.

Lischeid, G., H. Webber, M. Sommer, C. Nendel, and F. Ewert, 2022: Machine learning in crop yield modelling: A powerful tool, but no surrogate for science. *Agric. For. Meteor.*, **312**, 108698, https://doi.org/10.1016/j.agrformet.2021.108698.

Liu, Q., M. Yang, K. Mohammadi, D. Song, J. Bi, and G. Wang, 2022: Machine learning crop yield models based on meteorological features and comparison with a process-based model. *Artif. Intell. Earth Syst.*, **1**, e220002, https://doi.org/10.1175/AIES-D-22-0002.1.

Mamalakis, A., E. A. Barnes, and I. Ebert-Uphoff, 2022: Investigating the fidelity of explainable artificial intelligence methods for applications of convolutional neural networks in geoscience. *Artif. Intell. Earth Syst.*, **1**, e220012, https://doi.org/10.1175/AIES-D-22-0012.1.

——, ——, and ——, 2023: Carefully choose the baseline: Lessons learned from applying XAI attribution methods for regression tasks in geoscience. *Artif. Intell. Earth Syst.*, **2**, e220058, https://doi.org/10.1175/AIES-D-22-0058.1.

Martínez-Ferrer, L., M. Piles, and G. Camps-Valls, 2020: Crop yield estimation and interpretability with Gaussian processes. *IEEE Geosci. Remote Sens. Lett.*, **18**, 2043–2047, https://doi.org/10.1109/LGRS.2020.3016140.

Mateo-Sanchis, A., M. Piles, J. Amorós-López, J. Muñoz Marí, J. E. Adsuara, A. Moreno-Martínez, and G. Camps-Valls, 2021: Learning main drivers of crop progress and failure in Europe with interpretable machine learning. *Int. J. Appl. Earth Obs. Geoinf.*, **104**, 102574, https://doi.org/10.1016/j.jag.2021.102574.

——, J. E. Adsuara, M. Piles, J. Munoz-Marí, A. Perez-Suay, and G. Camps-Valls, 2023: Interpretable long short-term memory networks for crop yield estimation. *IEEE Geosci. Remote Sens. Lett.*, **20**, 2501105, https://doi.org/10.1109/LGRS.2023.3244064.

Meyer, H., and E. Pebesma, 2021: Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol. Evol.*, **12**, 1620–1633, https://doi.org/10.1111/2041-210X.13650.

——, and ——, 2022: Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nat. Commun.*, **13**, 2208, https://doi.org/10.1038/s41467-022-29838-9.

——, C. Reudenbach, T. Hengl, M. Katurji, and T. Nauss, 2018: Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environ. Modell. Software*, **101**, 1–9, https://doi.org/10.1016/j.envsoft.2017.12.001.

——, ——, S. Wöllauer, and T. Nauss, 2019: Importance of spatial predictor variable selection in machine learning applications—Moving from data reproduction to spatial prediction. *Ecol. Modell.*, **411**, 108815, https://doi.org/10.1016/j.ecolmodel.2019.108815.

Molnar, C., and Coauthors, 2022: General pitfalls of model-agnostic interpretation methods for machine learning models. *xxAI—Beyond Explainable AI*, A. Holzinger et al., Eds., Lecture Notes in Computer Science, Vol. 13200, Springer, 39–68, https://doi.org/10.1007/978-3-031-04083-2_4.

Müller, C., and Coauthors, 2017: Global gridded crop model evaluation: Benchmarking, skills, deficiencies and implications. *Geosci. Model Dev.*, **10**, 1403–1422, https://doi.org/10.5194/gmd-10-1403-2017.

——, and Coauthors, 2019: The Global Gridded Crop Model Intercomparison phase 1 simulation dataset. *Sci. Data*, **6**, 50, https://doi.org/10.1038/s41597-019-0023-8.

——, and Coauthors, 2021: Exploring uncertainties in global crop yield projections in a large ensemble of crop models and CMIP5 and CMIP6 climate scenarios. *Environ. Res. Lett.*, **16**, 034040, https://doi.org/10.1088/1748-9326/abd8fc.

Newman, S. J., and R. T. Furbank, 2021: Explainable machine learning models of major crop traits from satellite-monitored continent-wide field trial data. *Nat. Plants*, **7**, 1354–1363, https://doi.org/10.1038/s41477-021-01001-0.

Oliveira, S., F. Oehler, J. San-Miguel-Ayanz, A. Camia, and J. M. C. Pereira, 2012: Modeling spatial patterns of fire occurrence in

Mediterranean Europe using multiple regression and random forest. *For. Ecol. Manage.*, **275**, 117–129, https://doi.org/10.1016/j.foreco.2012.03.003.

Ortiz-Bobea, A., H. Wang, C. M. Carrillo, and T. R. Ault, 2019: Unpacking the climatic drivers of US agricultural yields. *Environ. Res. Lett.*, **14**, 064003, https://doi.org/10.1088/1748-9326/ab1e75.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Peichl, M., S. Thober, L. Samaniego, B. Hansjürgens, and A. Marx, 2021: Machine-learning methods to assess the effects of a non-linear damage spectrum taking into account soil moisture on winter wheat yields in Germany. *Hydrol. Earth Syst. Sci.*, **25**, 6523–6545, https://doi.org/10.5194/hess-25-6523-2021.

Ploton, P., and Coauthors, 2020: Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.*, **11**, 4540, https://doi.org/10.1038/s41467-020-18321-y.

Portmann, F. T., S. Siebert, and P. Döll, 2010: MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochem. Cycles*, **24**, GB1011, https://doi.org/10.1029/2008GB003435.

Pudil, P., J. Novovičová, and J. Kittler, 1994: Floating search methods in feature selection. *Pattern Recognit. Lett.*, **15**, 1119–1125, https://doi.org/10.1016/0167-8655(94)90127-9.

Ribeiro, A. F. S., A. Russo, C. M. Gouveia, P. Páscoa, and J. Zscheischler, 2020: Risk of crop failure due to compound dry and hot extremes estimated with nested copulas. *Biogeosciences*, **17**, 4815–4830, https://doi.org/10.5194/bg-17-4815-2020.

Richards, J., R. Huser, E. Bevacqua, and J. Zscheischler, 2023: Insights into the drivers and spatio-temporal trends of extreme Mediterranean wildfires with statistical deep-learning. *Artif. Intell. Earth Syst.*, https://doi.org/10.1175/AIES-D-22-0095.1, in press.

Roberts, D. R., and Coauthors, 2017: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, **40**, 913–929, https://doi.org/10.1111/ecog.02881.

Romano, Y., M. Sesia, and E. Candès, 2020: Deep knockoffs. *J. Amer. Stat. Assoc.*, **115**, 1861–1872, https://doi.org/10.1080/01621459.2019.1660174.

Ruane, A. C., and Coauthors, 2021: Strong regional influence of climatic forcing datasets on global crop model ensembles. *Agric. For. Meteor.*, **300**, 108313, https://doi.org/10.1016/j.agrformet.2020.108313.

Rudin, C., 2019: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, **1**, 206–215, https://doi.org/10.1038/s42256-019-0048-x.

Ryo, M., 2022: Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. *Artif. Intell. Agric.*, **6**, 257–265, https://doi.org/10.1016/j.aiia.2022.11.003.

Schaphoff, S., U. Heyder, S. Ostberg, D. Gerten, J. Heinke, and W. Lucht, 2013: Contribution of permafrost soils to the global carbon budget. *Environ. Res. Lett.*, **8**, 014026, https://doi.org/10.1088/1748-9326/8/1/014026.

Schauberger, B., S. Rolinski, and C. Müller, 2016: A network-based approach for semi-quantitative knowledge mining and its application to yield variability. *Environ. Res. Lett.*, **11**, 123001, https://doi.org/10.1088/1748-9326/11/12/123001.

Schierhorn, F., M. Hofmann, T. Gagalyuk, I. Ostapchuk, and D. Müller, 2021: Machine learning reveals complex effects of climatic means and weather extremes on wheat yields during different plant developmental stages. *Climatic Change*, **169**, 39, https://doi.org/10.1007/s10584-021-03272-0.

Schmidt, L., F. Heße, S. Attinger, and R. Kumar, 2020: Challenges in applying machine learning models for hydrological inference: A case study for flooding events across Germany. *Water Resour. Res.*, **56**, e2019WR025924, https://doi.org/10.1029/2019WR025924.

Seneviratne, S., and Coauthors, 2021: Weather and climate extreme events in a changing climate. *Climate Change 2021: The Physical Science Basis*, V. Masson-Delmotte et al., Eds., Cambridge University Press, 1513–1766, https://doi.org/10.1017/9781009157896.013.

Sheffield, J., G. Goteti, and E. F. Wood, 2006: Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling. *J. Climate*, **19**, 3088–3111, https://doi.org/10.1175/JCLI3790.1.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, 2008: Conditional variable importance for random forests. *BMC Bioinf.*, **9**, 307, https://doi.org/10.1186/1471-2105-9-307.

van der Wiel, K., F. M. Selten, R. Bintanja, R. Blackport, and J. A. Screen, 2020: Ensemble climate-impact modelling: Extreme impacts from moderate meteorological conditions. *Environ. Res. Lett.*, **15**, 034050, https://doi.org/10.1088/1748-9326/ab7668.

Vishwakarma, S., X. Zhang, and V. Lyubchich, 2022: Wheat trade tends to happen between countries with contrasting extreme weather stress and synchronous yield variation. *Commun. Earth Environ.*, **3**, 261, https://doi.org/10.1038/s43247-022-00591-7.

Vogel, E., M. G. Donat, L. V. Alexander, M. Meinshausen, D. K. Ray, D. Karoly, N. Meinshausen, and K. Frieler, 2019: The effects of climate extremes on global agricultural yields. *Environ. Res. Lett.*, **14**, 054010, https://doi.org/10.1088/1748-9326/ab154b.

Vogel, J., and Coauthors, 2021: Identifying meteorological drivers of extreme impacts: An application to simulated crop yields. *Earth Syst. Dyn.*, **12**, 151–172, https://doi.org/10.5194/esd-12-151-2021.

Vorndran, M., A. Schütz, J. Bendix, and B. Thies, 2022: Current training and validation weaknesses in classification-based radiation fog nowcast using machine learning algorithms. *Artif. Intell. Earth Syst.*, **1**, e210006, https://doi.org/10.1175/AIES-D-21-0006.1.

Waha, K., L. G. J. van Bussel, C. Müller, and A. Bondeau, 2012: Climate-driven simulation of global crop sowing dates. *Global Ecol. Biogeogr.*, **21**, 247–259, https://doi.org/10.1111/j.1466-8238.2011.00678.x.

Webber, H., G. Lischeid, M. Sommer, R. Finger, C. Nendel, T. Gaiser, and F. Ewert, 2020: No perfect storm for crop yield failure in Germany. *Environ. Res. Lett.*, **15**, 104012, https://doi.org/10.1088/1748-9326/aba2a4.

Wolanin, A., G. Mateo-García, G. Camps-Valls, L. Gómez-Chova, M. Meroni, G. Duveiller, Y. Liangzhi, and L. Guanter, 2020: Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt. *Environ. Res. Lett.*, **15**, 024019, https://doi.org/10.1088/1748-9326/ab68ac.

Zscheischler, J., and Coauthors, 2020: A typology of compound weather and climate events. *Nat. Rev. Earth Environ.*, **1**, 333–347, https://doi.org/10.1038/s43017-020-0060-z.