Environmental Data Science (2023), 2: e27, 1–13
doi:10.1017/eds.2023.21

CAMBRIDGE UNIVERSITY PRESS

METHODS PAPER

Selecting robust features for machine-learning applications using multidata causal discovery

Saranya Ganesh S.[1], Tom Beucler[1], Frederick Iat-Hin Tam[1], Milton S. Gomez[1], Jakob Runge[2,3] and Andreas Gerhardus[2]

[1]Institute of Earth Surface Dynamics, University of Lausanne (UNIL), Lausanne, Switzerland
[2]Institute of Data Science, German Aerospace Center (DLR), Jena, Germany
[3]Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, Berlin, Germany
Corresponding author: Saranya Ganesh S; Email: saranyaganesh.s@gmail.com

Received: 04 May 2023; **Accepted:** 31 May 2023

Keywords: causal feature selection; machine learning; multivariate time series analysis; tropical cyclones

Abstract

Robust feature selection is vital for creating reliable and interpretable machine-learning (ML) models. When designing statistical prediction models in cases where domain knowledge is limited and underlying interactions are unknown, choosing the optimal set of features is often difficult. To mitigate this issue, we introduce a multidata (M) causal feature selection approach that simultaneously processes an ensemble of time series datasets and produces a single set of causal drivers. This approach uses the causal discovery algorithms PC_1 or PCMCI that are implemented in the Tigramite Python package. These algorithms utilize conditional independence tests to infer parts of the causal graph. Our causal feature selection approach filters out causally spurious links before passing the remaining causal features as inputs to ML models (multiple linear regression and random forest) that predict the targets. We apply our framework to the statistical intensity prediction of Western Pacific tropical cyclones (TCs), for which it is often difficult to accurately choose drivers and their dimensionality reduction (time lags, vertical levels, and area-averaging). Using more stringent significance thresholds in the conditional independence tests helps eliminate spurious causal relationships, thus helping the ML model generalize better to unseen TC cases. M-PC_1 with a reduced number of features outperforms M-PCMCI, noncausal ML, and other feature selection methods (lagged correlation and random), even slightly outperforming feature selection based on explainable artificial intelligence. The optimal causal drivers obtained from our causal feature selection help improve our understanding of underlying relationships and suggest new potential drivers of TC intensification.

Impact Statement

While causal feature selection helps design more robust ML models, its joint application to multiple datasets remains limited because standard causal discovery algorithms output a different set of drivers for each dataset, which is impractical. To mitigate this issue, we apply a newly developed "multidata" causal feature selection approach, which identifies a single set of optimal causal drivers from an ensemble of multivariate time series. Applied to the statistical prediction of TC intensity, our approach outperforms standard feature selection methods by helping simple regression algorithms better generalize to unseen cases. In addition to making our models

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

robust, causal feature selection also eliminates redundant predictors while identifying new ones, leading to lighter models and aiding scientific discovery.

1. Introduction

Machine learning (ML) combines statistical methods and numerical optimization to learn a group of tasks from data. Progress in computational capabilities, combined with the availability of large amounts of data, allows the development of ML models to predict and understand nonlinear systems such as climate processes and extreme weather events. For environmental applications, processing big data that are nonlinearly related often requires (a) dimensionality reduction; and (b) strategically selecting the model's features to make ML models cheaper to run, generalize better, and easier to explain (Guyon and Elisseeff, 2003; Yu et al., 2020, 2022). This article compares different methods to discover a subset of the most relevant features in environmental datasets (Guyon and Elisseeff, 2003; Post et al., 2016; Li et al., 2017) and explores the effect of causal feature selection on statistical prediction skill. For this, we work at the intersection of causal inference and ML, an active area of research (Chen et al., 2020) because causal relations help acquire robust knowledge beyond the support of observed data distributions (Schölkopf et al., 2021). Causal inference can broadly be categorized into three research directions: (a) causal representation learning; (b) causal discovery; and (c) causal reasoning (Schölkopf et al., 2021; Kaddour et al., 2022). To select features, we here explore the use of *causal discovery*, a methodology for learning qualitative cause-and-effect relationships between a collection of variables from data that have not been obtained under controlled experimental conditions (Spirtes et al., 2000; Peters et al., 2017). Incorporating causal relationships in ML models via feature selection can make ML models more interpretable (Guyon and Elisseeff, 2003; Runge et al., 2015b; Yu et al., 2022; Iglesias-Suarez et al., 2023) and less susceptible to overfitting (Aliferis et al., 2010a, 2010b; Runge et al., 2015a).

There are two main challenges when applying causal discovery in environmental sciences. The first challenge is algorithmic: Often environmental data consists of multiple realizations of the same process with slight differences, and causal discovery algorithms that apply to such multiple realization problems remain underexploited (Yu et al., 2020). The second challenge is the lack of benchmarking: Causal feature selection is rarely compared against other feature selection methods for ML-based predictions. Here, we address these two gaps by introducing a causal feature selection framework to estimate causal relationships from multiple time series datasets (Runge et al., 2015a, 2015b, 2019a, 2019b; Yu et al., 2019; Runge, 2020). We compare feature selection algorithms by training simple ML prediction models for each of the selected sets of features and evaluating their predictive performances. Our framework is applied to the prediction of tropical cyclone (TC) intensity to demonstrate that causal feature selection (a) improves the out-of-sample skill, and (b) uncovers the best predictors in real-world situations.

2. Methodology: Causal Feature Selection for Multiple Realizations

Our implementation of causal feature selection, Geiger et al. (1990), Pena et al. (2007), and Gao and Ji (2017) uses the recently developed multidata (M) functionality for two causal discovery algorithms based on time-series, explained below. Our *multidata* causal discovery approach used to preselect causally relevant predictors has two steps: (a) the causal discovery algorithms; and (b) applying these algorithms to a dataset comprising data from multiple sources. From a causal perspective, the setup used in this study is simplified because only the variables that are time-lagged with respect to the target variables are considered potential predictors. As a result, causal discovery effectively reduces to a feature selection algorithm that removes all those predictors which are (conditionally) independent of the target (given the other predictors) and which hence do not provide any additional information for predicting the target. The multidata functionality itself, however, is more general and also applies to the time series causal discovery tasks that also consider contemporaneous causal relationships.

https://doi.org/10.1017/eds.2023.21 Published online by Cambridge University Press

Here, we explore the use of the causal discovery algorithms PC_1 and PCMCI (Runge et al., 2019b) for causal feature selection. The PC_1 algorithm is a variant of the PC algorithm (Spirtes et al., 2000). First, PC_1 initializes the potential causal drivers $pa(Y_t)$ of each target variable Y_t as the set of all variables $pa(Y_t) = \left\{X^i_{t-\tau} | i = 1, \ldots, N_X, \tau = \tau_{min}, \ldots, \tau_{\max}\right\}$ within the considered range $[\tau_{min}, \tau_{max}]$ of time lags, where the X^i with $i = 1, \ldots, N_X$ are the potential predictors and where τ_{min} and τ_{max}, respectively, are the minimal and maximal time lags at which direct causal influences can occur. Then, PC_1 iteratively removes variables from $pa(Y_t)$ that are irrelevant or redundant for the prediction of Y_t. Specifically, PC_1 removes elements $X^i_{t-\tau}$ from $pa(Y_t)$ that are conditionally independent of Y_t given subsets $S_k \subseteq pa(Y_t)$ whose cardinality k increases iteratively: For $k = 0$ all $X^i_{t-\tau}$ with $X^i_{t-\tau} \perp\!\!\!\perp Y_t$ are removed, for $k = 1$ those with $X^i_{t-\tau} \perp\!\!\!\perp Y_t | S_1$, where S_1 is the strongest driver from the previous step, for $k = 2$ those with $X^i_{t-\tau} \perp\!\!\!\perp Y_t | S_2$, where S_2 are the two strongest drivers from the previous step, and so on. In this work, conditional independence is tested using partial correlation (in general, though, the algorithm can be combined with any conditional independence test). The corresponding independence test is based on a standard significance level $pc_\alpha = 0.02$ and uses a strength of association that is based on the absolute partial correlation value. This iteration is continued until k is greater than the cardinality of $pa(Y_t)$. The PC algorithm is different from PC_1 in so far as that PC, for every k, does not only test for conditional independence given exactly one cardinality k subset of $pa(Y_t)$ but tests for conditional independence given all cardinality k subsets of $pa(Y_t)$.

The PCMCI algorithm (Runge et al., 2019b), after first running PC_1, reinitializes all links and then subjects all links to the momentary conditional independence (MCI) tests $X^i_{t-\tau} \perp\!\!\!\perp Y_t | pa(Y_t) \setminus \{X^i_{t-\tau}\}, pa(X^i_{t-\tau})$, removing the link if independence is not rejected. Here, the condition on $pa\left(X^i_{t-\tau}\right)$ helps to remove false positives that tend to be inflated due to autocorrelation. Controlling false positives is important for a causal discovery setting but is not necessarily important for a causal feature selection/prediction setting as considered in this article. Within the study presented here, we employ both the PC_1 and the PCMCI algorithm to empirically analyze which of the two methods is preferable for causal feature selection. As with all causal discovery methods, PC_1 and PCMCI rely on certain assumptions. The essential assumption is that conditional independencies in the data are in one-to-one correspondence with *d*-separations (Pearl, 1988) in the causal graph (Geiger et al., 1990; Verma and Pearl, 1990; Spirtes et al., 2000; Pearl, 2009). Moreover, both methods assume *causal sufficiency* (Spirtes et al., 2000), that is, the absence of unobserved variables that causally influence two observed variables. The latter assumption is not necessarily fulfilled in our context, even though we included a range of potential predictors. This means that some of the obtained causal features might still be spurious and may not work if the target distribution differs from the training distribution (out-of-distribution prediction).

When PC_1 and PCMCI are applied to a *single* multivariate time series, samples are drawn from this time series in a sliding-window fashion. The drawn set is internally passed to the statistical hypothesis tests of conditional independencies. For this sliding-window approach to be valid, the causal relationships need to remain unchanged throughout the time series (*causal stationarity* assumption). When PC_1 and PCMCI are applied to *multiple* multivariate time series, if all time series of this ensemble can be assumed to share the same causal relationships within specific time ranges, then we can combine the sample sets from all ensemble members[1] into a single, larger dataset. This larger dataset, which includes data from multiple sources (e.g., from multiple storms), is then internally passed to the conditional independence tests. The PC_1 and PCMCI algorithms can then proceed as usual. Consequently, although the input is an ensemble of multivariate time series, the output is still a single set of predictors. In addition to its practicality, our *multidata* approach benefits from an enlarged set of samples, increasing the power of the conditional independence tests. Hence, we expect the sets of predictors obtained by running multidata causal discovery on an ensemble of time series to be more reliable than the sets of predictors obtained by running causal discovery on any single member of this ensemble—if the assumption of a common causal

[1] Each of the sample sets is obtained in a sliding window fashion from one of the member time series.

https://doi.org/10.1017/eds.2023.21 Published online by Cambridge University Press

structure holds. An alternative approach would be to run PC_1 and PCMCI on any single member time series and then appropriately aggregate the resulting sets of predictors (across the members).

3. Application: Statistical Prediction of Tropical Cyclone Intensity

3.1. Motivation

The increasing frequency of intense TCs (Emanuel, 2005; Knutson et al., 2020) combined with growing coastal populations have escalated the vulnerability of the tropical urban coasts. For context, more than half of Earth's population is projected to live in the tropics by 2050 (Edelman et al., 2014) and more than a billion people worldwide could be living in low-elevation coastal zones by 2060, particularly in South and East Asia (Neumann et al., 2015). Predicting storm intensity changes, including rapid intensification in TCs, remains a major challenge (DeMaria et al., 2014), because of unresolved complexities of storm dynamics in numerical models. Furthermore, numerical models suffer from a reduction in forecast skills with an increase in lead time (Ganesh et al., 2018). An alternative to numerical forecasting is statistical forecasting, as statistical models can improve cyclogenesis and intensity forecasts (Kim et al., 2019; Chen et al., 2020). For instance, statistical models based on logistic regression, random forest, decision tree, and randomized trees (Su et al., 2020) outperformed the National Hurricane Center in predicting TC rapid intensifications over the Atlantic and Eastern Pacific basins (Kaplan et al., 2010; Rozoff and Kossin, 2011). A potential drawback of statistical models is that it is often difficult to choose appropriate predictors for reliable forecasts. To better predict TC intensity, the models need to represent the physical mechanisms behind TC intensification more accurately; these include large-scale circulations, local conditions, and internal processes (Kaplan et al., 2010; Emanuel and Zhang, 2016). We argue that one way to make statistical models more robust is to apply causal discovery algorithms and eliminate causally spurious predictors. In this study, we apply the PC_1 and PCMCI methods to generate a single set of causally relevant predictors from multiple TC time series.

3.2. Data

The TC dataset is created using multiple environmental variables at different pressure levels known to be favorable for TC intensification (Sikora, 1976; Petty and Hobgood, 2000; Li et al., 2011) from the global high-resolution ECMWF ReAnalysis-5 (ERA5; Hersbach et al., 2020) with 25 km horizontal resolution, and 3-hourly temporal resolution (see Section A of the Supplementary Material for the full list). Here, we selected a total of 260 TC cases spanning from 2001 to 2020 in the Northwest Pacific basin (WPAC). The TCs with a lifetime of more than 6 days up to landfall are selected for the study to understand the effect of environmental parameters on TC intensity up to 3 days time lag, so each case has a time span from genesis up to landfall based on each TC best track[2]. Rather than directly feeding the time series of predictor variables for the cases at each grid point around the storm, the values are averaged in horizontal areas defined with respect to the TC Center. Each atmospheric predictor is post-processed into two sets of time series representing inner-core (TC center to a radius of 200 km) and outer-core characteristics (annulus from a radius of 200–800 km). The choice of averaged areas follows the current practice in operational statistical intensity prediction schemes (DeMaria et al., 2005). The distinction between outer-core and inner-core processes is justified because TC intensity is affected by environmental conditions in the storm's neighborhood and internal processes within the storm (Sitkowski and Barnes, 2009; Hendricks et al., 2019). From an ML perspective, this preliminary dimensionality reduction removes features with high spatial correlations, reduces the complexity of the statistical models, and possibly improves model generalizability by removing some of the predictors' spatial heterogeneity in different storms.

Once this preliminary dimensionality reduction is done, our goal is to eliminate spurious *features*, here defined based on meteorological variable, vertical (pressure) levels, time lag, and horizontal averaging sector (inner or outer core). We describe each TC using a total of $N_X = 234$ time series of horizontally

[2] TC tracks are obtained from the International Best Track Archive for Climate Stewardship (Knapp et al., 2010).

https://doi.org/10.1017/eds.2023.21 Published online by Cambridge University Press

Table 1. *R^2 score for each experiment's best model on the validation set, along with the number of selected features (in parentheses)*

ML models/target		Training (no. of features)			Validation			Test		
		P_{min} (hPa)	V10 (ms^{-1})	Precip $\times 10^{-3}$ (km^2)	P_{min}	V10	Precip	P_{min}	V10	Precip
Causal-RF		0.93 (26)	0.89 (17)	0.83 (123)	0.87	0.81	0.65	0.88	0.78	0.62
Causal-MLR		**0.87** (17)	**0.84** (31)	**0.71** (90)	**0.88**	**0.82**	**0.68**	**0.89**	**0.80**	**0.62**
Noncausal-RF	All	0.93 (3,978)	0.88 (3,978)	0.75 (3,978)	0.77	0.74	0.65	0.89	0.79	0.58
	Lagged	0.96 (480)	0.93 (560)	0.79 (80)	0.85	0.81	0.69	0.89	0.81	0.61
	Random	0.96 (870)	0.93 (770)	0.85 (970)	0.79	0.77	0.59	0.87	0.77	0.56
Noncausal-MLR	All	0.99 (3,978)	0.98 (3,978)	0.96 (3,978)	−0.94	−10.85	−127.98	0.51	−0.01	−0.39
	Lagged	0.92 (440)	0.64 (40)	0.68 (120)	0.84	0.54	0.65	0.92	0.59	0.64
	Random	0.91 (420)	0.83 (130)	0.69 (290)	0.78	0.76	0.62	0.86	0.75	0.54
	XAI	0.92 (240)	0.88 (420)	0.70 (140)	0.84	0.82	0.69	0.91	0.79	0.63
LSTM		0.87 (3,978)	0.81 (3,978)	0.71 (3,978)	0.77	0.75	0.65	0.81	0.75	0.61

Note: *Causal-MLR* (bold) gives the best performance with the least features.

averaged 3D variables at given vertical levels and 2D variables (see Supplementary Table 1 for details). With regards to the time lags, we explore the time steps between 24 hr before the target (corresponding to $\tau_{min} = 8$) and 72 hr before the target (corresponding to $\tau_{max} = 24$). This results in a total number of 3,978 (234 potential predictors × 17 time steps) for the causal algorithms, which eliminate the spurious links between the features and the targets. We randomly split the data *by TC* to avoid spatiotemporal correlation: Out of the selected 260 TCs, we randomly split 205 cases from 2001 to 2020 into a training set (150 cases) and validation set (55 cases) while keeping 55 cases from recent years (2017–2020) in the test set, without any overlaps. The regression task is to forecast three variables with a 1-day lead time, including (1) maximum wind speed at 10 m (max. 10 m wind, in m/s), (2) minimum sea-level pressure (MSLP, in hPa), and (3) horizontally integrated total precipitation (Tot. Intg. Precip. in km^2) accumulated over 3-hourly intervals. Maximum sustained wind speed at 10 m (averaged over 1 min, 3 min, or 10 min depending on the Regional Specialized Meteorological Centre) is the standard measurement for the intensity currently used operationally. We include MSLP as it correlates better with TC damage (Atkinson and Holliday, 1977; Kaplan et al., 2010). Additionally, MSLP is easier to estimate as it is an integrated quantity and only requires a couple of measurements near the storm center. Finally, we included total integrated precipitation as a potential target because most fatalities and damage from TCs are caused by heavy precipitation and storm surges (Rappaport, 2014).

3.3. Causal machine learning

In this section, we describe the feature selection methodology in the context of TC intensity prediction. Our causal feature selection framework is shown in Figure 1. Once the four-dimensional fields have been reduced to time series, we align the time series in the training set based on the minimum pressure value recorded during each TC's lifetime, which is a smoother measure of TC intensity than maximum surface wind speeds (Chavas et al., 2017). Temporally aligning the multivariate time series of different ensemble members is key, as the resulting ensemble is more likely to satisfy the common causal structure assumption, improving prediction skills using causal feature selection. After aligning the time series, we feed the training set as inputs to the PC_1 and PCMCI algorithms (both implemented in Tigramite) to

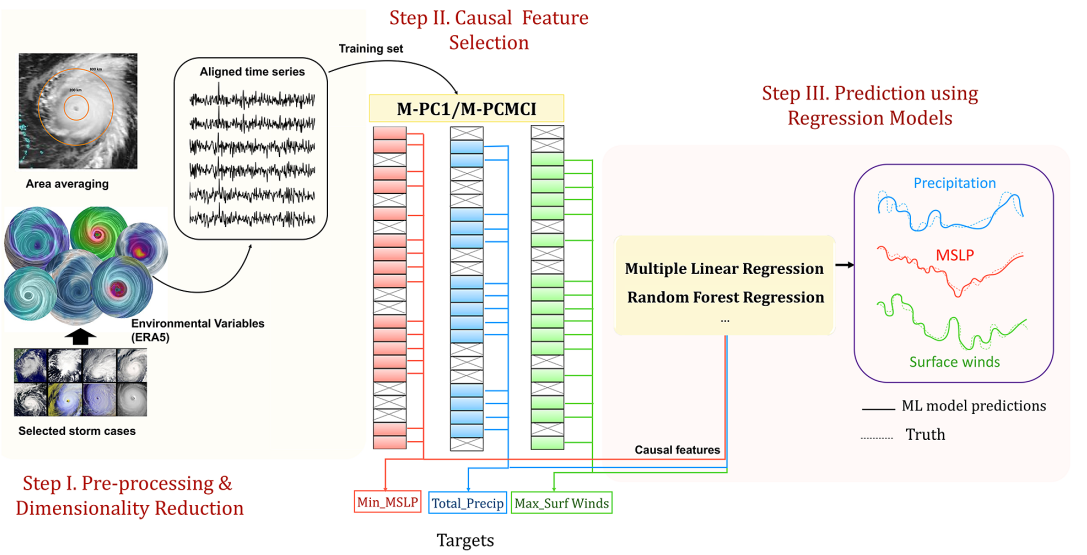

Figure 1. *Multidata causal feature selection applied to TC prediction: After reducing the dimensionality of spatiotemporal fields to yield time series for several TC cases (Step I), the ensemble of aligned time series is fed to the multidata causal discovery algorithm to calculate the optimal set of causal drivers (Step II), which can be fed to a regression algorithm to make robust predictions (Step III).*

https://doi.org/10.1017/eds.2023.21 Published online by Cambridge University Press

extract the most significant causal features. Here, an input feature may be defined as an environmental or derived variable (see Supplementary Table 1) at any given pressure level which is spatially averaged either by the inner or outer core radii at a given 3-hourly time-step. We stress that PC_1 and PCMCI are only applied to the training data. The implementation of both PC_1 and PCMCI contains several hyperparameters, including minimum and maximum time lags for the analysis, fixed to 1 day (8 timesteps) and 3 days (24 timesteps), respectively, for our prediction task. Further tunable hyperparameters are the significance level for conditional independence testing (pc_α) and a significance threshold for the p-value matrix (alpha-level, only used for PCMCI), which control the selected number of features.

Once PC_1 and PCMCI have selected the most significant causal drivers of the targets from the ensemble of time series, these drivers are used as inputs to the prediction model. We logarithmically vary the values of pc_α and alpha-level, which in effect controls the number of selected features, as more stringent (pc_α) values will result in the selection of fewer and more significant features. From this set of experiments, we determine the best hyperparameters suitable for each target of interest by maximizing the validation performance of the trained regression models. We use multiple linear regression (MLR) and random forest (RF) regression models to predict the targets from the causally selected features. The MLR algorithms for causal-MLR experiments were prepared using the Scikit-Learn (Pedregosa et al., 2011) implementation of the linear regression algorithm and its corresponding default parameters. Each MLR algorithm was trained to predict one of three unscaled target variables using the selected, standard-scaled features. We also included RF regression models using the same causal feature selection set-up (Figure 1) to explore the impact of causal feature selection for more complex nonlinear regression methods. We used the RF regressor from the sci-kit-learn library (Pedregosa et al., 2011) to prepare the causal-RF and optimized its hyperparameters with a randomized search.

3.4. Noncausal machine-learning baselines

This study is motivated by the working hypothesis that regression models using causally selected features outperform noncausal baselines in terms of generalization. Here, *noncausal baselines* subsume both the case of no feature selection and the case of feature selection based on noncausal criteria. We compare our causal feature selection to noncausal feature selection methods such as lagged correlation, random selection as well as an explainable artificial intelligence (XAI) method of feature selection using RF regression (more details are provided in Section C of the Supplementary Material). To test our causal approach's ability to effectively use time lags, we also train a long short-term memory (LSTM) network

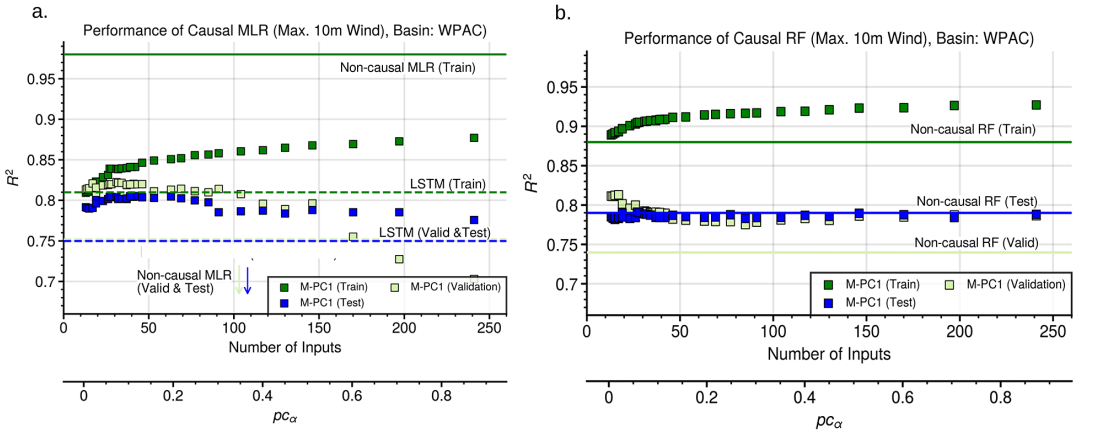

Figure 2. *(a) Causal-MLR models using M-PC_1 systematically outperform LSTMs (dashed line) on all sets and their noncausal counterparts (solid lines) on the validation and test sets. (b) Causal-RF models outperform their noncausal counterparts (solid lines) on the training and validation sets.*

https://doi.org/10.1017/eds.2023.21 Published online by Cambridge University Press

using all time lags between τ_{min} and τ_{max} and without feature selection. We implement the LSTM using the PyTorch (Paszke et al., 2019) library and conducted a hyperparameter search with the Optuna (Akiba et al., 2019) framework. A more detailed description of the architecture is provided in Section C of the Supplementary Material.

4. Results

4.1. Performance of causal machine learning

Our first objective is to find the set of causal features that are best linked to the intensification in TCs at a lead time of 1-day. To measure the suitability of a set of causal features, we evaluate how MLR as well as RF models trained with the causally selected features perform when predicting TCs that are unseen during training. We evaluate prediction skill holistically (see Section D of the Supplementary Material), but focus on the coefficient of determination (R^2) in the main text, with $R^2 = 1$ corresponding to a perfect prediction and $R^2 = 0$ to an error of one standard deviation. In Figure 2, we show the performance of Causal-ML models to predict the maximum winds, 24 hr in advance using M-PC_1 method. A less stringent significance threshold results in a larger set of features that are retained during training, which has a clear negative effect on the model validation performance. We found that feature selection using M-PCMCI is comparable to M-PC_1 when the minimum time lag is 6 hr (shown in Section B of the Supplementary Material), but the performance of M-PCMCI drastically deteriorates when we increase the minimum time lag to 1 day. Here, we only show the causal ML results based on M-PC_1 here. For comparison, similar experiments with reduced lead times where minimum and maximum time lags are set to 6 hr and 2 days, respectively, are shown in Section B of the Supplementary Material. PCMCI's main advantage is to better control false positives in the presence of strong autocorrelation (Runge et al., 2019b), which is more important for an actual causal discovery setting than for the causal prediction setting considered in this article.

Causal-MLR scores are better than those of noncausal MLRs (Figure 2a), which use all inputs without feature selection. The noncausal baselines clearly overfit the training set. The causal MLR is also compared with a recurrent neural network using an LSTM layer, and causal MLRs outperform the best LSTM model for all targets, which is remarkable given their simplicity. When comparing the RF models, we find that causal-RF scores are better than noncausal-RF for the validation set, whereas test set scores are comparable for wind speed predictions. In general, the (R^2) values are highest for the predictions of MSLP, followed by wind and total integrated precipitation (Figure 2 and Supplementary Figures 1 and 2). The optimal set of causal drivers that performs well on the training and validation sets (without leading to overfitting) is sparse, containing only 31 features in the causal-MLR case for predicting maximum wind (Table 1). This result suggests that many of the features are spuriously linked to TC intensity and can be removed without sacrificing the predictive skill of simple MLR models compared to a noncausal-RF baseline. Similar results are obtained for the prediction of other targets, as shown in Supplementary Figures 1 and 2.

4.2. Comparison with feature selection baselines

Our second objective is to compare the performance of causal MLRs to MLRs with noncausal feature selection baselines (described in Section C of the Supplementary Material). For the maximum wind predictions (Figure 3), PC_1 consistently outperforms the two simpler feature selection baselines (random and lagged correlation), especially on the validation and test sets (Figure 3b,c). Lagged correlation, in particular, selects sets of predictors that perform very poorly in comparison, especially during validation. The ability of an XAI-based feature selection to capture nonlinearities seems to improve predictor selection, resulting in R^2 values that are almost comparable to the PC_1 causal feature selection method. Nevertheless, the causal PC_1 method retains an advantage for very sparse models (less than 50 input features), suggesting that the initial selection of causally relevant predictors allows these sparse models to beat the corresponding noncausal sparse models. PC_1 performs better than the other methods for the two

https://doi.org/10.1017/eds.2023.21 Published online by Cambridge University Press

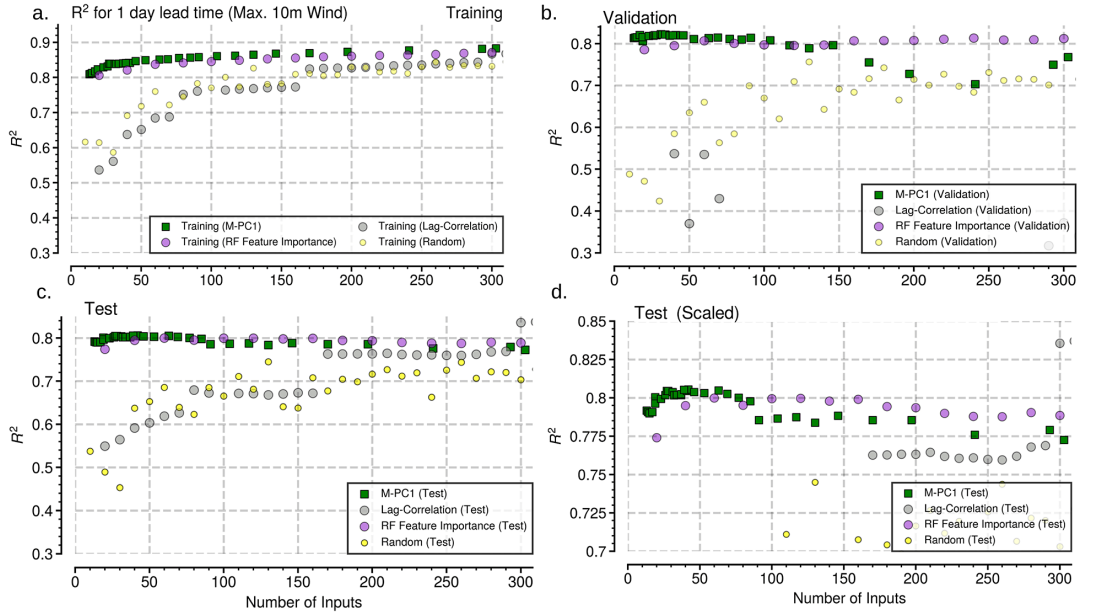

Figure 3. *While (a) both causal and noncausal models fit the training set better when their number of features is increased, M-PC*$_1$ *causal feature selection provides the best generalization to unseen cases in the validation (b) and test sets (c and d zoomed-in version), especially when the number of input features is below 100 (d). For all methods, selected features are fed to MLR for predicting maximum winds for WPAC TCs at a lead time of 1 day.*

other targets, which is shown in Section C of the Supplementary Material (Supplementary Figures 3–6). We note that lagged correlation performance is comparable to causal-MLR and XAI-based feature selection in predicting total integrated precipitation. This motivates adapting our conditional independence tests for non-normal distributions, which we leave for future work. The performance comparison based on (R^2) of all best ML models used in the study, along with their number of input features, is listed in Table 1.

4.3. Optimal causal features

Here, we expand upon our results from the previous section to understand *why* causal-MLR models outperform the models that use other feature selection methods. For this purpose, we rely on the frequency of predictor selection (across the models) by the best-performing causal-MLR and lagged correlation MLR models.[3] We find that both methods choose different predictors for the maximum wind predictions while identifying inner core relative humidity as critical for wind prediction. However, divergence is a major predictor in the causal-MLR (Figure 4a), despite not being in the 10 most frequently selected predictors for the lagged-correlation models (Figure 4d). The vertical distribution (Figure 4b,e) and the time lag information (Figure 4c,f) of the most frequently chosen features reveal several differences in the causal models as compared to the lagged correlation models. Unlike the lag correlation method, the causal method selects features at specific vertical levels and time lags that are most informative to the prediction (Figure 4b) rather than placing importance over a wide range of vertical levels (Figure 4e) and lags. The PC$_1$ algorithm iteratively removes variables from the parent set $pa(Y_t)$ that are irrelevant or redundant for

[3] Best causal-MLR models are defined as model with R^2 within 1% of the best validation R^2. This threshold is relaxed to 10% for lagged correlation models to sample a comparable number of features.

https://doi.org/10.1017/eds.2023.21 Published online by Cambridge University Press

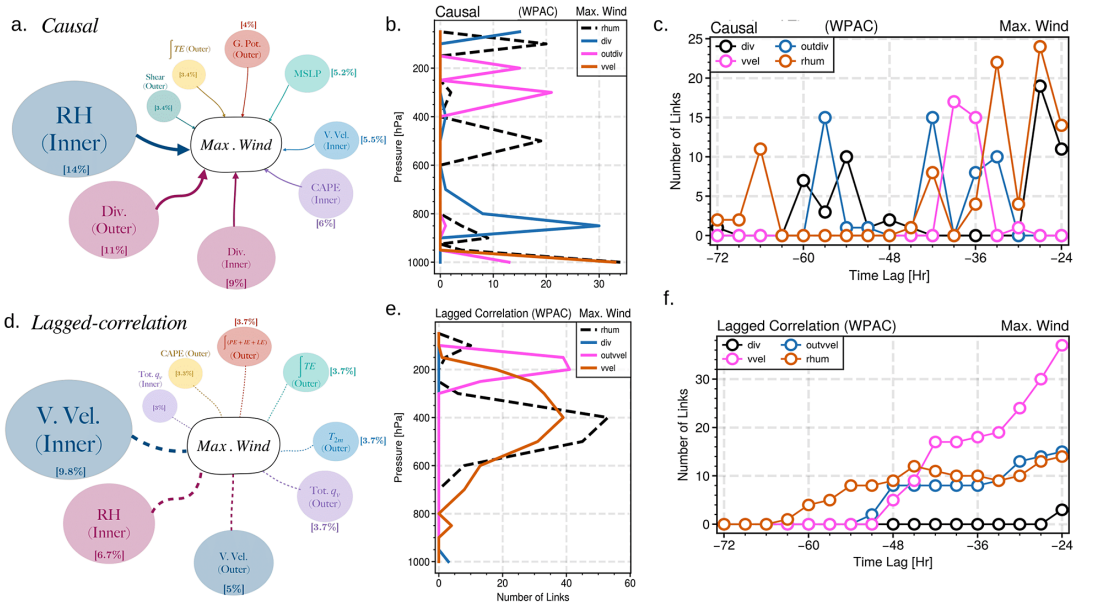

Figure 4. *Most frequently and significant predictors used by the best causal-MLR model organized by (a) top nine meteorological variables; (b) pressure level; and (c) time lag. (d–f) Most frequently selected features for the lag correlation method. For the two rightmost columns, we retained the four most frequent features (Relative humidity (inner), Vertical velocity (inner) and horizontal divergence (inner and outer).*

the prediction of the target, Y_t via conditional independence tests (here, based on partial correlation). PC_1 then ranks features based on significance test statistics, which gives a good measure of predictor relevance. Once X_t is in the parent set $pa(Y_t)$, its neighbors will be iteratively removed because they are not conditionally independent of X_t. This confirms the interpretation that the selected time lags and vertical levels are "most predictive" of Y_t, and that the spatiotemporal neighbors of X_t are eliminated because of redundancy, which is due to the high spatiotemporal correlations in our dataset. Next, from a scientific viewpoint, the causal models clearly emphasize the low-level inner-core convergence (div), middle and upper tropospheric relative humidity (RH, rhum)in the inner core and the upper-level divergence (outdiv) in the outer core as most important predictors whereas the lagged correlation models rely on middle- and upper-tropospheric vertical motions (vvel) for the prediction task. Finally, in the time-lag plots (Figure 4c,f), the divergence links in the causal models are chosen at time lags of more than 2 days (-60 hr $< t < -50$ hr), while lagged correlation models focus on features at the shortest lead times ($t > -48$ hr) as they are more correlated with decreasing time lags.

Causal-MLR models rely on low-level convergence and upper-level divergence at longer time lags. In contrast, the lagged correlation MLR models mostly rely on mid- to upper-tropospheric vertical motion at shorter lead times. One way to interpret this is that the mid-tropospheric vertical motion could be a confounder, which is removed by the PC_1 algorithm. In this case, the difference in generalization skill may be attributed to the lagged correlation MLRs making predictions based on causally spurious links. The causal relationship involved here can be understood in mass adjustment terms: mass conservation requires low-level convergence and upper-level divergence to be balanced by upward mass transport. This upward motion can invigorate convection and aid TC intensification. Hence, the vertical motion should be considered as a *consequence* of divergence rather than an independent process that drives TC intensification *by itself*. We believe that the removal of mid-level vertical motion in the PC_1 features shows that causal discovery algorithms can successfully remove causally spurious links.

https://doi.org/10.1017/eds.2023.21 Published online by Cambridge University Press

5. Conclusion

This article described a causal feature selection framework to predict and understand complex geophysical events that can be considered multiple realizations of the same process with small perturbations. We applied this framework to multiple TC time series to identify common causal links and used them as predictors in MLR and RF regression models. Our results show that causal feature selection is superior to traditional feature selection methods for finding sets of predictors that help regression models generalize to unseen TC cases, especially for very sparse models (Figure 3). Of the two causal methods, we find that the PC_1 algorithm is more appropriate for feature selection, as it only keeps the most informative features, effectively removes confounding features (e.g., mid-tropospheric vertical motion in Figure 4), and is less sensitive to the forecast lead time (Supplementary Figures 3–5) than PCMCI. Temporally aligning the multiple time series based on a common reference point before causal feature selection tangibly improves model prediction skills. The retention of spurious links in the lag correlation models negatively affected generalizability. From these observations, we conclude that causal feature selection holds potential in our continued effort to improve statistical TC intensity models. Future efforts will involve (a) assessing whether current operational intensity prediction baselines can be improved by the causality-based predictor selection; (b) expanding the study to all ocean basins; and (c) discovering new potential predictors that may improve operational TC intensity predictions.

While not studied in this article, the multidata causal discovery also opens the possibility to analyze systems whose causal structure changes in time: If one can align the individual member time series on a common time axis and can assume that, although changing in time, their causal structures are the same, then a dataset for independence testing can still be created by taking one sample per member time series. Repeating this procedure for every time step would yield one set of predictors per variable and per time step. Similarly, one could obtain one set of predictors per variable and per time window in a sequence of time windows (useful for slowly varying causal structures). Finally, we note that the multidata approach does not rely on any particular causal discovery algorithm. Therefore, while not shown here, the multidata approach can also be employed with the $PCMCI^+$ algorithm (Runge, 2020), a variant of PCMCI that allows contemporaneous causal influences and the latent-PCMCI (LPCMCI) algorithm (Gerhardus and Runge, 2020), a variant of PCMCI that allows for contemporaneous causal influences and latent confounders (available within the open-source Python package Tigramite). Lastly, one could further optimize predictions by selecting the subset of causal predictors with the highest (validation set-)skill as discussed from an information-theoretic perspective in Runge et al. (2015a). In our context, however, iterating through all feature subsets is computationally prohibitive.

Acknowledgments. We thank the DCSR at UNIL for providing the computational resources and technical support.

Author contribution. Conceptualization: S.G.S.; T.B.; F.I.T., M.G.; Data curation: S.G.S.; F.I.T.; T.B.; Data visualization: S.G.S., F.I.T.; T.B.; Methodology: J.R.; A.G.; T.B.; S.G.S.; Writing original draft: S.G.S. All authors contributed to writing and review. All authors approved the revised manuscript.

Competing interest. The authors declare that no competing interests exist.

Data availability statement. The codes and tutorials for multidata causal discovery are freely available in the Tigramite GitHub repository and have been archived in Zenodo at https://doi.org/10.5281/zenodo.7747255. Sample code for the application are available at Causal-ML GitHub repository and have been archived in Zenodo at https://doi.org/10.5281/zenodo.7907217. The WPAC TC data are from the IBtrACS data archive. ERA5 datasets were downloaded from the Copernicus website (multiple pressure levels as well as single pressure levels).

Ethics statement. The research meets all ethical guidelines.

Funding statement. This research was supported by the canton of Vaud in Switzerland. J.R. has received funding from the European Research Council (ERC) Starting Grant CausalEarth under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 948112).

Provenance statement. This article is part of the Climate Informatics 2023 proceedings and was accepted in *Environmental Data Science* on the basis of the Climate Informatics peer review process.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/eds.2023.21.

References

Akiba T, Sano S, Yanase T, Ohta T and Koyama M (2019) Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*. New York, USA: Association for Computing Machinery, pp. 2623–2631.

Aliferis CF, Statnikov A, Tsamardinos I, Mani S and Koutsoukos XD (2010a) Local causal and Markov blanket induction for causal discovery and feature selection for classification part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research 11*(1), 171–234.

Aliferis CF, Statnikov A, Tsamardinos I, Mani S and Koutsoukos XD (2010b) Local causal and Markov blanket induction for causal discovery and feature selection for classification part II: Analysis and extensions. *Journal of Machine Learning Research 11*(1), 235–284.

Atkinson GD and Holliday CR (1977) Tropical cyclone minimum sea level pressure/maximum sustained wind relationship for the western North Pacific. *Monthly Weather Review 105*(4), 421–427.

Chavas DR, Reed KA and Knaff JA (2017) Physical understanding of the tropical cyclone wind-pressure relationship. *Nature Communications 8*(1), 1360.

Chen H, Harinen T, Lee JY, Yung M and Zhao Z (2020) CausalML: Python package for causal machine learning. Preprint, arXiv: 2002.11631v2

Chen R, Zhang W and Wang X (2020) Machine learning in tropical cyclone forecast modeling: A review. *Atmosphere 11*(7), 676.

DeMaria M, Mainelli M, Shay LK, Knaff JA and Kaplan J (2005) Further improvements to the statistical hurricane intensity prediction scheme (SHIPS). *Weather and Forecasting 20*(4), 531–543.

DeMaria M, Sampson CR, Knaff JA and Musgrave KD (2014) Is tropical cyclone intensity guidance improving? *Bulletin of the American Meteorological Society 95*(3), 387–398.

Edelman A, Gelding A, Konovalov E, McComiskie R, Penny A, Roberts N, Trewin DJ, Ziembicki M, Trewin B, Cortlet R, Hemingway J, Isaac J and Turton S (2014) *State of the Tropics 2014 Report*. Cairns: James Cook University.

Emanuel K (2005) Increasing destructiveness of tropical cyclones over the past 30 years. *Nature 436*(7051), 686–688.

Emanuel K and Zhang F (2016) On the predictability and error sources of tropical cyclone intensity forecasts. *Journal of the Atmospheric Sciences 73*(9), 3739–3747.

Ganesh SS, Sahai AK, Abhilash S, Joseph S, Dey A, Mandal R, Chattopadhyay R and Phani R (2018) A new approach to improve the track prediction of tropical cyclones over North Indian Ocean. *Geophysical Research Letters 45*(15), 7781–7789.

Gao T and Ji Q (2017) Efficient Markov blanket discovery and its application. *IEEE Transactions on Cybernetics 47*(5), 1169–1179.

Geiger D, Verma T and Pearl J (1990) Identifying independence in Bayesian networks. *Networks 20*(5), 507–534.

Gerhardus A and Runge J (2020) High-recall causal discovery for autocorrelated time series with latent confounders. *Networks 33*, 12615–12625.

Guyon I and Elisseeff A (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research 3*, 1157–1182.

Hendricks EA, Braun SA, Vigh JL and Courtney JB (2019) A summary of research advances on tropical cyclone intensity change from 2014–2018. *Tropical Cyclone Research and Review 8*(4), 219–225.

Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, Simmons A, Soci C, Abdalla S, Abellan X, Balsamo G, Bechtold P, Biavati G, Bidlot J, Bonavita M, Chiara G, Dahlgren P, Dee D, Diamantakis M, Dragani R, Flemming J, Forbes R, Fuentes M, Geer A, Haimberger L, Healy S, Hogan RJ, Hólm E, Janisková M, Keeley S, Laloyaux P, Lopez P, Lupu C, Radnoti G, Rosnay P, Rozum Ia, Vamborg Fn, Villaume Sd and Thépaut JN (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society 146* (730), 1999–2049.

Iglesias-Suarez F, Gentine P, Solino-Fernandez B, Beucler T, Pritchard M, Runge J and Eyring V (2023) Causally-informed deep learning to improve climate models and projections. Preprint, arXiv preprint 2304.12952.

Kaddour J, Lynch A, Liu Q, Kusner MJ and Silva R (2022) Causal machine learning: A survey and open problems. Preprint, arXiv preprint arXiv 2206(15475).

Kaplan J, DeMaria M and Knaff JA (2010) A revised tropical cyclone rapid intensification index for the Atlantic and eastern North Pacific basins. *Weather and Forecasting 25*(1), 220–241.

Kim M, Park MS, Im J, Park S and Lee MI (2019) Machine learning approaches for detecting tropical cyclone formation using satellite data. *Remote Sensing 11*(10), 1195.

Knapp KR, Kruk MC, Levinson DH, Diamond HJ and Neumann CJ (2010) The international best track archive for climate stewardship (IBTrACS) unifying tropical cyclone data. *Bulletin of the American Meteorological Society 91*(3), 363–376.

Knutson T, Camargo SJ, Chan JC, Emanuel K, Ho CH, Kossin J, Mohapatra M, Satoh M, Sugi M, Walsh K and Wu L (2020) Tropical cyclones and climate change assessment: Part II: Projected response to anthropogenic warming. *Bulletin of the American Meteorological Society 101*(3), E303–E322.

Li G, **Ren B**, **Yang C and Zheng J** (2011) Revisiting the trend of the tropical and subtropical Pacific surface latent heat flux during 1977–2006. *JGR: Atmospheres 116*(D10), D10115 1–9.

Li J, **Cheng K**, **Wang S**, **Morstatter F**, **Trevino RP**, **Tang J and Liu H** (2017) Feature selection: A data perspective. *ACM Computing Surveys (CSUR) 50*(6), 1–45.

Neumann B, **Vafeidis AT**, **Zimmermann J and Nicholls RJ** (2015) Future coastal population growth and exposure to sea-level rise and coastal flooding-a global assessment. *PLoS One 10*(3), e0118571.

Paszke A, **Gross S**, **Massa F**, **Lerer A**, **Bradbury J**, **Chanan G**, **Killeen T**, **Lin Z**, **Gimelshein N**, **Antiga L and Desmaison A** (2019) Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, NIPS-2019, Vancouver, BC, Canada, pp. 1–12.

Pearl J (1988) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo: Morgan Kaufmann.

Pearl J (2009) Causal inference in statistics: An overview. *Statistical Surveys 3*, 96–146.

Pedregosa F, **Varoquaux G**, **Gramfort A**, **Michel V**, **Thirion B**, **Grisel O**, **Blondel M**, **Prettenhofer P**, **Weiss R**, **Dubourg V and Vanderplas J** (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12*, 2825–2830.

Pena JM, **Nilsson R**, **Björkegren J and Tegnér J** (2007) Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning 45*(2), 211–232.

Peters J, **Janzing D and Schölkopf B** (2017) *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: The MIT Press.

Petty KR and Hobgood JS (2000) Improving tropical cyclone intensity guidance in the eastern North Pacific. *Weather and Forecasting 15*(2), 233–244.

Post MJ, **Putten PVD and Rijn JNV** (2016) Does feature selection improve classification? A large scale experiment in OpenML. In *International Symposium on Intelligent Data Analysis*. Cham: Springer, pp. 158–170.

Rappaport EN (2014) Fatalities in the United States from Atlantic tropical cyclones: New data and interpretation. *Bulletin of the American Meteorological Society 95*(3), 341–346.

Rozoff CM and Kossin JP (2011) New probabilistic forecast models for the prediction of tropical cyclone rapid intensification. *Weather and Forecasting 26*(5), 677–689.

Runge J (2020) Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. *Proceedings of Machine Learning Research 124*, 1388–1397.

Runge J, **Donner RV and Kurths J** (2015a) Optimal model-free prediction from multivariate time series. *Physical Review E 91*, 052909.

Runge J, **Petoukhov V**, **Donges JF**, **Hlinka J**, **Jajcay N**, **Vejmelka M**, **Hartman D**, **Marwan N**, **Paluš M and Kurths J** (2015b) Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications 6*(1), 1–10.

Runge J, **Bathiany S**, **Bollt E**, **Camps-Valls G**, **Coumou D**, **Deyle E**, **Glymour C**, **Kretschmer M**, **Mahecha MD**, **Muñoz-Marí J**, **van Nes EH**, **Peters J**, **Quax R**, **Reichstein M**, **Scheffer M**, **Schölkopf B**, **Spirtes P**, **Sugihara G**, **Sun J**, **Zhang K and Zscheischler J** (2019a) Inferring causation from time series in earth system sciences. *Nature Communications 10*(1), 1–13.

Runge J, **Nowack P**, **Kretschmer M**, **Flaxman S and Sejdinovic D** (2019b) Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances 5*(11), eaau4996.

Schölkopf B, **Locatello F**, **Bauer S**, **Ke NR**, **Kalchbrenner N**, **Goyal A and Bengio Y** (2021) Toward causal representation learning. *Proceedings of the IEEE 109*(5), 612–634.

Sikora CR (1976) *An Investigation of Equivalent Potential Temperature as a Measure of Tropical Cyclone Intensity*. San Francisco: FLEET WEA. CENTRAL/JTWC FPO.

Sitkowski M and Barnes GM (2009) Low-level thermodynamic, kinematic, and reflectivity fields of hurricane Guillermo (1997) during rapid intensification. *Monthly Weather Review 137*(2), 645–663.

Spirtes P, **Glymour CN**, **Scheines R and Heckerman D** (2000) *Causation, Prediction, and Search*. Cambridge: The MIT Press.

Su H, **Wu L**, **Jiang JH**, **Pai R**, **Liu A**, **Zhai AJ**, **Tavallali P and DeMaria M** (2020) Applying satellite observations of tropical cyclone internal structures to rapid intensification forecast with machine learning. *Geophysical Research Letters 47*(17), e2020GL089102.

Verma T and Pearl J (1990) Causal networks: Semantics and expressiveness. *Machine Intelligence and Pattern Recognition 9*, 69–76.

Yu K, **Guo X**, **Liu L**, **Li J**, **Wang H**, **Ling Z and Wu X** (2020) Causality-based feature selection: Methods and evaluations. *ACM Computing Surveys 53*(5), 1–36.

Yu K, **Liu L**, **Li J**, **Ding W and Le TD** (2019) Multi-source causal feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 42*(9), 2240–2256.

Yu K, **Yang Y and Ding W** (2022) Causal feature selection with missing data. *ACM Transactions on Knowledge Discovery from Data 16*(4), 1–24.

Cite this article: S. S, Beucler T, Tam FI-H., Gomez MS, Runge J and Gerhardus A (2023). Selecting robust features for machine-learning applications using multidata causal discovery. *Environmental Data Science*, 2: e27. doi:10.1017/eds.2023.21