Time series causal relationships discovery through feature importance and ensemble models

Manuel Castro (castroavila@ic.unicamp.br)
Artificial Intelligence Lab., Recod.ai, Institute of Computing, University of Campinas (Unicamp), 13083-852 Campinas, SP

Pedro Ribeiro Mendes Júnior
Artificial Intelligence Lab., Recod.ai, Institute of Computing, University of Campinas (Unicamp), 13083-852 Campinas, SP

Aurea Soriano-Vargas
Artificial Intelligence Lab., Recod.ai, Institute of Computing, University of Campinas (Unicamp), 13083-852 Campinas, SP

Rafael De Oliveira Werneck
Artificial Intelligence Lab., Recod.ai, Institute of Computing, University of Campinas (Unicamp), 13083-852 Campinas, SP

Maiara Gonçalves
Center for Petroleum Engineering (CEPETRO), University of Campinas (Unicamp), 13083-970, Campinas, SP

Leopoldo Lusquino Filho
Group of Automation and Integrated Systems, São Paulo State University (Unesp), 18087-180, Sorocaba, SP

Renato Moura
Artificial Intelligence Lab., Recod.ai, Institute of Computing, University of Campinas (Unicamp), 13083-852 Campinas, SP

Marcelo Zampieri
Center for Petroleum Engineering (CEPETRO), University of Campinas (Unicamp), 13083-970, Campinas, SP

Oscar Linares
Artificial Intelligence Lab., Recod.ai, Institute of Computing, University of Campinas (Unicamp), 13083-852 Campinas, SP

Vitor Ferreira
Center for Petroleum Engineering (CEPETRO), University of Campinas (Unicamp), 13083-970, Campinas, SP

Alexandre Ferreira
Artificial Intelligence Lab., Recod.ai, Institute of Computing, University of Campinas (Unicamp), 13083-852 Campinas, SP

Alessandra Davólio
Center for Petroleum Engineering (CEPETRO), University of Campinas (Unicamp), 13083-970, Campinas, SP
Denis Schiozer
School of Mechanical Engineering (FEM), University of Campinas (Unicamp), 13083-970, Campinas, SP
Anderson Rocha
Artificial Intelligence Lab., Recod.ai, Institute of Computing, University of Campinas (Unicamp), 13083-852 Campinas, SP

Article

Keywords:

Posted Date: February 16th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2566176/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on July 14th, 2023. See the published version at https://doi.org/10.1038/s41598-023-37929-w.

Time series causal relationships discovery through feature importance and ensemble models

Manuel Castro[1, *], Pedro Ribeiro Mendes Júnior[1], Aurea Soriano-Vargas[1], Rafael de Oliveira Werneck[1], Maiara Gonçalves[2], Leopoldo Lusquino Filho[4], Renato Moura[1], Marcelo Zampieri[2], Oscar Linares[1], Vitor Ferreira[2], Alexandre Ferreira[1], Alessandra Davólio[2], Denis Schiozer[3], and Anderson Rocha[1]

[1]Artificial Intelligence Lab., Recod.ai, Institute of Computing, University of Campinas (Unicamp), 13083-852 Campinas, SP, Brazil.
[2]Center for Petroleum Engineering (CEPETRO), University of Campinas (Unicamp), 13083-970, Campinas, SP, Brazil.
[3]School of Mechanical Engineering (FEM), University of Campinas (Unicamp), 13083-970, Campinas, SP, Brazil.
[4]Group of Automation and Integrated Systems, São Paulo State University (Unesp), 18087-180, Sorocaba, SP, Brazil.
[*]castroavila@ic.unicamp.br

ABSTRACT

Inferring causal relationships from observational data is a key challenge when seeking to understand the interpretability of Machine Learning models. Given the ever increasing amount of observational data available in many areas, Machine Learning algorithms used for forecasting have increased their complexity, leading to a less understandable path of how a decision is made by the model. With this in mind, we propose leveraging ensemble models, e.g., Random Forest, to assess which input features the trained model pays more attention to when making a forecast and, in this way, establish causal relationships between the variables. The advantage of those algorithms is the possibility of obtaining the *feature importance*, which allows us to build the causal network. We apply our methods to estimate causality in time series for two domains: One climate dataset and two oil field production datasets. We aim to perform *causal discovery*, i.e., establish the existing connections between the variables in each dataset. Through an iterative process of improving the forecasting of a target's value, we evaluate whether the forecasting improves by adding information from a new potential driver; if so, we state that the driver causally affects the target. On the oil field related datasets, the results we obtained based on causal analysis agree with the interwell connections already confirmed by tracer information; for the cases when the tracer data are available, we used it as our ground truth. This agreement between both estimated and confirmed connections provides us the confidence about the effectiveness of our proposed methodology. To our knowledge this is the first time causal analysis using solely production data is employed to discover interwell connections in an oil field related dataset.

Introduction and motivation

Understanding how a set of variables are related is becoming an important aspect in different fields of science, e.g., social and behavioral sciences, neuroscience, and econometrics. For instance, understanding what the efficiency is of a drug in a population, if a public policy has the expected effects on society[1], how major climate effects such as El Niño Southern Oscillation (ENSO) influence remote regions, or through which pathways different regions of the brain interact[2] help comprehend the dynamics behind complex systems. To answer those questions, the standard procedure is to manipulate the value of a variable by performing a Randomized Controlled Trial (RCT). However, in many cases, it is not possible to perform such experiments because it is unethical, practically impossible, or expensive, such as performing controlled experiments to establish the causal relationship between smoking and lung cancer (see Pearl & Mackenzie[3], Chap. 5).

The advent of Machine Learning (ML) algorithms, the increasing amount of collected data, and the increased hardware processing power now available, allow processing those data and applying complex algorithms to find causal relationships among variables (i.e., establish causality). The classical definition of causality is that $\mathbf{X} \rightarrow \mathbf{Y}$ ($\rightarrow$ meaning a causal link) if and only if an intervention or manipulation in $\mathbf{X}$ affects $\mathbf{Y}$[4], where $\mathbf{X}$ is known as the driver and $\mathbf{Y}$ as the target. The first approach to measure the degree of associations among variables is to compute the correlation. However, the correlation is symmetrical and does not take into account the directionality of the relationship. On the other hand, causation is asymmetrical and provides the directionality of the relation among the variables[5].

Why do we need to establish causality and not only work on establishing statistical associations? One major reason is

that causality enables us to predict the underlying dynamics of the system, pinpointing which variables can be intervened to achieve the desired output[6] through a causal model. With this model, we can use what Pearl and Mackenzie in their book — *The book of why* [3, Chap. 1] — defined as **The ladder of Causation**: A 3-rung ladder that goes from simple data association to hypothetical imaging scenarios ("what if I had done"). In this paper, we deal with the first rung, establishing the associations between variables in a dataset; current ML algorithms are also on this rung since they only rely on data and are thus unable to operate in unseen scenarios. Once the associations are discovered, we can go to the second rung called *interventions*: Given a relationship between $\mathbf{X}$ and $\mathbf{Y}$, this level answers questions such as "what would be $\mathbf{Y}$ if I intervene $\mathbf{X}$, does $\mathbf{Y}$ happen?". Finally, the last rung is called *counterfactuals* and answers questions such as "Was it $\mathbf{X}$ that caused $\mathbf{Y}$?" or "What if $\mathbf{X}$ had not occurred?"[7].

Why is causality hard to establish? When attempting to assess causality, we must ensure that every possible variable that might be involved in the system's dynamic must be considered to avoid spurious correlations. In a real-world scenario, this condition is rarely accomplished: There might be unobserved *confounders* (when investigating a potential causal-and-effect relationship, a *confounding variable* is a third variable that influences both the supposed cause and the supposed effect), the data are scarce, and few variables may change simultaneously and influence the outcome we observe, either directly or indirectly. This lack of control over the system's dynamics makes it even harder to claim that a causal relationship exists, which leads researchers to depend highly on the opinion of experts in the area to have more confidence in the findings.

In this paper, we focus on time series (ordered sequences of values), in which the goal of causal discovery is to establish the causal links, including the time lags (time-lagged causal discovery). Lags tell us the time delay between a cause and the occurrence of its effect. Related to time series, ML algorithms have been used for classification[8], clustering[9], and forecasting[10,11], but little work has been done to employ ML to infer causal relationships in time series (see Moraffah et al.[12] for a full review in the subject).

We organize the paper into the following sections: Section **Time series causality assessment** details causality assessment; Section **Proposed methodology** describes our proposed methodology to assess causality in time series; Section **Applications** presents applications of causal discovery applied in two different domains — oil field production data and climate; finally, Section **Conclusions and future work** outlines final remarks discussing our main findings.

Time series causality assessment

In this section, we formalize the causality assessment.

The causal relationships

Given a dataset $\mathscr{X}$ with N time series of the same length T, i.e., $\mathscr{X} = \{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^N\} \in \mathbb{R}^{N \times T}$, we want to discover the causal relationships between all N time series in $\mathscr{X}$ and the time lag between a cause and the occurrence of its effect[5]. Figure 1 shows a sketch of how the process of causal discovery in time series works: From a set $\mathscr{X}$ of variables, possible causal links are tested from variable $\mathbf{X}^l$ to $\mathbf{X}^k$, $k,l = 1,2,\ldots,N$ (including links from the past of $\mathbf{X}^k$ itself, i.e., $l = k$). This allows us to reconstruct the underlying causal dependencies (solid black arrows), discard spurious associations (red arrows; when the correlation is significant, but there is no causal dependence), and the time dependence in the system because of the lags τ. The main challenges faced by any method that seeks causal discovery are:

1. Distinguish direct from indirect causes. In Figure 1, variable $\mathbf{X}^1$ directly affects both $\mathbf{X}^0$ and $\mathbf{X}^3$, but $\mathbf{X}^2$ has an indirect effect on $\mathbf{X}^0$ through $\mathbf{X}^1$ ($\mathbf{X}^1$ acts as a mediator).

2. Distinguish instantaneous causal effects, i.e., $\tau = 0$.

3. Deal with *confounders*: In Figure 1, $\mathbf{X}^1$ is a common cause for both $\mathbf{X}^0$ and $\mathbf{X}^3$, which explains the spurious correlation between $\mathbf{X}^0$ and $\mathbf{X}^3$.

Related work

Because of the vast amount of time series data in different areas, several efforts have been made to tackle causality in this type of data. Over the last decades, different methods have been proposed to assess this challenge. Those methods can be categorized as follows[5,12]:

Methods based on Granger causality

One of the first approaches to tackle causality in time series was suggested by Granger et al[13]. Given two time series, $\mathbf{X}$ (possible driver) and $\mathbf{Y}$ (target), it says $\mathbf{X}$ Granger-cause $\mathbf{Y}$, if and only if considering the past of $\mathbf{X}$, improves the forecasting

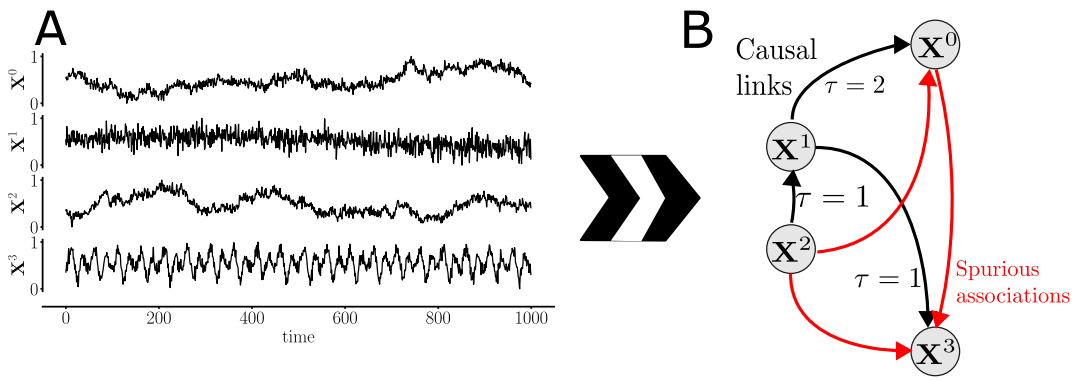

Figure 1. Given a set of multivariate time series (A), causal discovery aims to estimate the underlying causal dependencies, including the time lags τ (labels on the edges) (B). Spurious associations can appear due to either common drivers, e.g., $\mathbf{X}^0 \leftarrow \mathbf{X}^1 \rightarrow \mathbf{X}^3$ (correlation between $\mathbf{X}^0$ and $\mathbf{X}^3$) or indirect paths, e.g., $\mathbf{X}^2 \rightarrow \mathbf{X}^1 \rightarrow \mathbf{X}^0$ (correlation between $\mathbf{X}^2$ and $\mathbf{X}^0$) or $\mathbf{X}^2 \rightarrow \mathbf{X}^1 \rightarrow \mathbf{X}^3$ (correlation between $\mathbf{X}^2$ and $\mathbf{X}^3$). Adapted from[2].

of $\mathbf{Y}$ at time t. It means that $\mathbf{X}$ contains unique information about $\mathbf{Y}$, not contained in $\mathbf{Y}$'s past. Mathematically, this can be expressed as a Vector Autoregressive Models (VAR):

$$\mathbf{Y}_t = \sum_{\tau=1}^{\tau_{\max}} a_\tau \mathbf{Y}_{t-\tau} + \sum_{\tau=1}^{\tau_{\max}} b_\tau \mathbf{X}_{t-\tau} + \eta_t, \tag{1}$$

in which $\tau_{\max}$ is the maximum lag into the past that we want to consider.

If $\mathbf{X}$ Granger-cause $\mathbf{Y}$, it means some values b_τ are not zero. The main drawback of this approach is that it assumes a linear relationship among the variables, and it might not be suitable for every case in which we want to establish causal relationships. Some extensions of Granger causality applied to multivariate time series can be found in Haufe et al.[14] and Siggiridou & Kugiumtzis[15].

Conditional Independence-based methods

Those methods test conditional independence relations between variables and their past, i.e., it checks for conditional independence between $\mathbf{X}^k_{t-\tau}$ and $\mathbf{X}^l_t$ given the past of those variables (time series). Some assumptions are made[2,4]:

Time-order Cause precedes effects.

Causal sufficiency All direct common drivers are observed.

Causal Markov Condition If two variables are not connected in the causal graph conditioned on their parents (see Spirtes et al.[16], Chap. 3), they are conditionally independent.

Faithfulness If two variables are independent given a subset of variables, then they are not connected by a causal link in the graph.

Applications of this class of methods can be found in Chu & Glymour[17] and Runge[18].

Structural Equation-based models

Structural Equation Models (SEM) is a collection of statistical techniques to examine relationships between one or more independent variables and dependent variables[19]. Mathematically, it can be expressed as $\mathbf{X} := f(\mathbf{V}, \varepsilon_\mathbf{X})$, i.e., where each substantive variable $\mathbf{X}$ is a function of other variables $\mathbf{V}$, and a unique error term $\varepsilon_\mathbf{X}$[20]. The usage of the assignment operator (:=), rather than an equality operator, is because the equations must be interpreted causally: Manipulating a variable V can lead to a change in X. The substantive variables are the ones of interest but are not necessarily observed in their totality. An application using SEM applied to time series causal discovery, called Time Series Models with Independent Noise (TiMINo), can be found in Peters et al[21].

Deep Learning-based models

Traditional time series causal discovery methods assume some linearity in the time series, like Granger causality. However, real-world cases can be nonlinear, so those methods are unsuitable for such scenarios. To overcome those issues and take into account nonlinearity, some approaches have been proposed to model time series using neural network architectures like Multi

Later Perceptron (MLP), Recurrent Neural Network (RNN)[22] and Convolutional Neural Networks (CNNs)[5]. Usually, the inputs to those Deep Learning (DL)-based frameworks are the past lags of every time series (drivers), and the outputs are the future values of a single time series (target). CNNs use an attention mechanism to provide interpretability in the model. The attention mechanism coefficients can be interpreted as the model feature importance[23].

A summary of common causal discovery methods applied to time series is shown in Nauta et al.[5] Each method includes a description on whether it deals with aspects like *confounders*, *hidden confounders* and the necessity of stationary data, among other aspects. The main drawbacks of those methods relate to the existence of *hidden confounders* in the data as well as the need for stationary time series, since the methods could not work as expected when fed with this kind of data.

The main advantages of our approach to assessing causality (see Section **Proposed methodology**) are that:

i) It can look at causal relationships in nonlinear data;

ii) Handle multivariate and continuous time series;

iii) Take care of *confounders* (see Section **Step-by-step toy example**);

iv) Leverage the *feature importance* given by the ensemble models to pinpoint the delay (lag) between cause and effect.

We applied our method in datasets (both simulated and real) from different areas to test the applicability of our approach.

Proposed methodology

The general idea to assess causality is to be able to somehow pinpoint which predictors (variables and their lags) the algorithm considers the most when predicting an output, for instance, through the coefficients per lag in the Granger causality approach (see Section **Methods based on Granger causality**) or the usage of attention mechanisms in DL-related algorithms (see Section **Deep Learning-based models**). This section presents our proposed methodology, which we call Aleph, in honor of the important Latin American writer Jorge Luis Borges, who saw in the causal flow the foundation of both the world and the narrative activity, and who architected many of the short stories of his masterpiece, "The Aleph", from this perspective[24].

In our proposed methodology, we first define which variables (possible drivers) causally affect a target. Once a set of $\mathscr{D}$ drivers known to causally affect the target is discovered (Section **Time series causal discovery**), we proceed with the quantification of feature significance (Section **Assessing feature significance**), i.e., the discovery of the lags of influence from every driver to the target. For the first step, any regression method could be employed but, for the second one, we need ML algorithms that quantify the importance of each feature when predicting the output. Ensemble algorithms like Random Forest (RF) or CatBoost provide this information on feature importance. Therefore, we employ them to identify which predictors the model pays attention to the most when generating the forecasting and, from that information, we reconstruct the causal path (see Figure 1).

The general pipeline we propose to perform causal discovery is shown in Figure 2. The pipeline comprises two major steps:

i) Establish a baseline by considering only the target's past values (block `I`) to predict the target's current value and

ii) Using information from potential drivers in an attempt to improve the forecasting (block `II`).

The pipeline works as follows (and following the blocks identified by capital letters in Figure 2):

`(A)` Given a set of variables on which we want to perform causal discovery, we select one to be considered as a potential target, i.e., the one to which causal relationships will be assessed.

`(B)` Using past data from the potential target chosen in step `(A)`, we define feature vectors with past-lagged values from the potential target (see Figure 3).

`(C)` An ensemble ML model is trained based on the features defined in step `(B)` and used to predict the target's current value (more details in Section **Time series causal discovery**; see Figure 3).

`(D)` An initial *baseline* is established as the evaluation of a metric on the forecasting of step `(C)` in comparison with the ground-truth.

`(E)` A *set of potential drivers* that might causally affect the potential target is defined.

`(F)` If the *set of potential drivers* defined in step `(E)` is not empty, a driver is selected for consideration and removed from that set. Otherwise, the result of the pipeline is the *set of discovered drivers* incremented through step `(J)`.

(G) Using past data from the potential driver chosen in step (F), we define feature vectors with past-lagged values from the potential driver.

(H) Feature vectors from the potential target of step (B), from already *discovered drivers* through step (J), and from the potential driver under consideration chosen in step (F), are used to predict the target's current value (see Figure 3).

(I) The metric is evaluated based on the forecasting of the target's current value by the model in step (H) and compared against the *baseline*. If the forecasting of the model of step (H) is better than the *baseline*, we go through step (J); otherwise we go through step (K). In either case, we return to step (F).

(J) We state the driver selected in step (F) causally affects the target under evaluation, then the *baseline* is updated with the metric evaluated in step (I) and the driver under consideration is added to the *set of discovered drivers*.

(K) We consider that the driver under consideration does not causally affect the target, so we ignore it.

In-depth details are shown in Sections **Time series causal discovery**, **Assessing driver significance** and **Assessing feature significance**. A toy example was devised and each step taken to perform causal discovery (following the pipeline in Figure 2) is shown in Section **Step-by-step toy example**.

Time series causal discovery

Numerous ML techniques have been employed to determine causality in time series[5,25,26]. The idea behind those approaches is similar to the one of Granger causality, to test whether adding information from the past of possible drivers improves the forecasting of future values of the target. The advantage of ML models is that they are not restricted to variables with linear relationships (as some cases of Granger causality).

We employed two models for regression based on ensembles: RF[27] and CatBoost[28]. Both models are employed here as regressors for predicting continuous values. We first used them to evaluate if a variable significantly aids the regression in terms of performance, therefore considering that variable as the target variable's driver. Given the variables $\mathbf{X}$ and $\mathbf{Y}$, to determine if $\mathbf{X}$ drives (causally affects) $\mathbf{Y}$, we define the past values of both variables as features $\mathbf{X}_{t-\tau_{\max}},\ldots,\mathbf{X}_{t-1},\mathbf{Y}_{t-\tau_{\max}},\ldots,\mathbf{Y}_{t-1}$, having $\mathbf{Y}_t$ as target, as depicted in **Driver-evaluation experiment** of Figure 3. Similarly, we proceed to build features from $\mathbf{Y}$ as well with feature vectors $\mathbf{Y}_{t-\tau_{\max}},\ldots,\mathbf{Y}_{t-1}$ also with target $\mathbf{Y}_t$, as depicted in **Baseline experiment** of Figure 3. In any case, the goal is to predict $\mathbf{Y}$'s current value $\mathbf{Y}_t$ by using that past information. If information from $\mathbf{X}$ (lagged values) improves the forecasting of $\mathbf{Y}_t$ compared to the baseline, and that improvement is *significant* according to the method described in Section **Assessing driver significance**, we consider that $\mathbf{X}$ causally affects $\mathbf{Y}$. Therefore, $\mathbf{X}$ is included in the list of *discovered drivers* before the evaluation of a new driver.

One might wonder if the order of consideration of drivers from the set of possible ones matters. As we will see, for the oil field production data, we considered the order of the drivers based on the geographical distance from the target (the closer, the first to be considered). For the climate dataset, we considered an arbitrary order based on the order of specification of the variables in the dataset. Although we did not evaluate the influence of the order in those datasets, we experimented with the synthetic case of Section **Step-by-step toy example**. In that case, there are eight possible drivers which lead to a permutation of 40320 possible orders of consideration. From those, we randomly tested 100 cases and observed that our method obtains the same ground truth set of drivers in most of them, indicating that the order does not seem to matter much.

Assessing driver significance

As we previously mentioned, it is not enough that a variable $\mathbf{X}$ improves the forecasting of $\mathbf{Y}_t$ to have it considered as a driver. The key point in our approach to assessing causality is if there is a *significant improvement* in predicting the target's value by adding information from potential drivers. To quantify the quality of the forecasting, we can use any metric commonly used for regression. To compute the significance, as we evaluate a new driver, we perform a random shuffling of the features coming from that driver. We shuffle the features, train a model, make a forecasting on the testing set, and retrieve the quality of the forecasting (metric), repeating the process n times. We end up having a set $\mathscr{M} = \{m_1, m_2, \ldots, m_n\}$ of values for the metric due to the shuffling process and, additionally, we have the metric value m_{real} we obtain when training and testing the model without shuffling the data. The goal is to assess the statistical significance of m_{real} against the distribution given by $\mathscr{M}$ at a significance level α. We fit the data in $\mathscr{M}$ with a statistical distribution and compute the probability of m_{real} being obtained by chance. If this probability is less than α, we consider the driver causally affecting the target.

Assessing feature significance

Similarly, we proceed to assess the significance of each feature. Once we establish the set of drivers $\mathscr{D} = \{\mathbf{X}^1, \mathbf{X}^2, \ldots, \mathbf{X}^D\}$ which causally affect the target, we shuffle each feature from the drivers in $\mathscr{D}$ individually, keeping the other features unchanged.

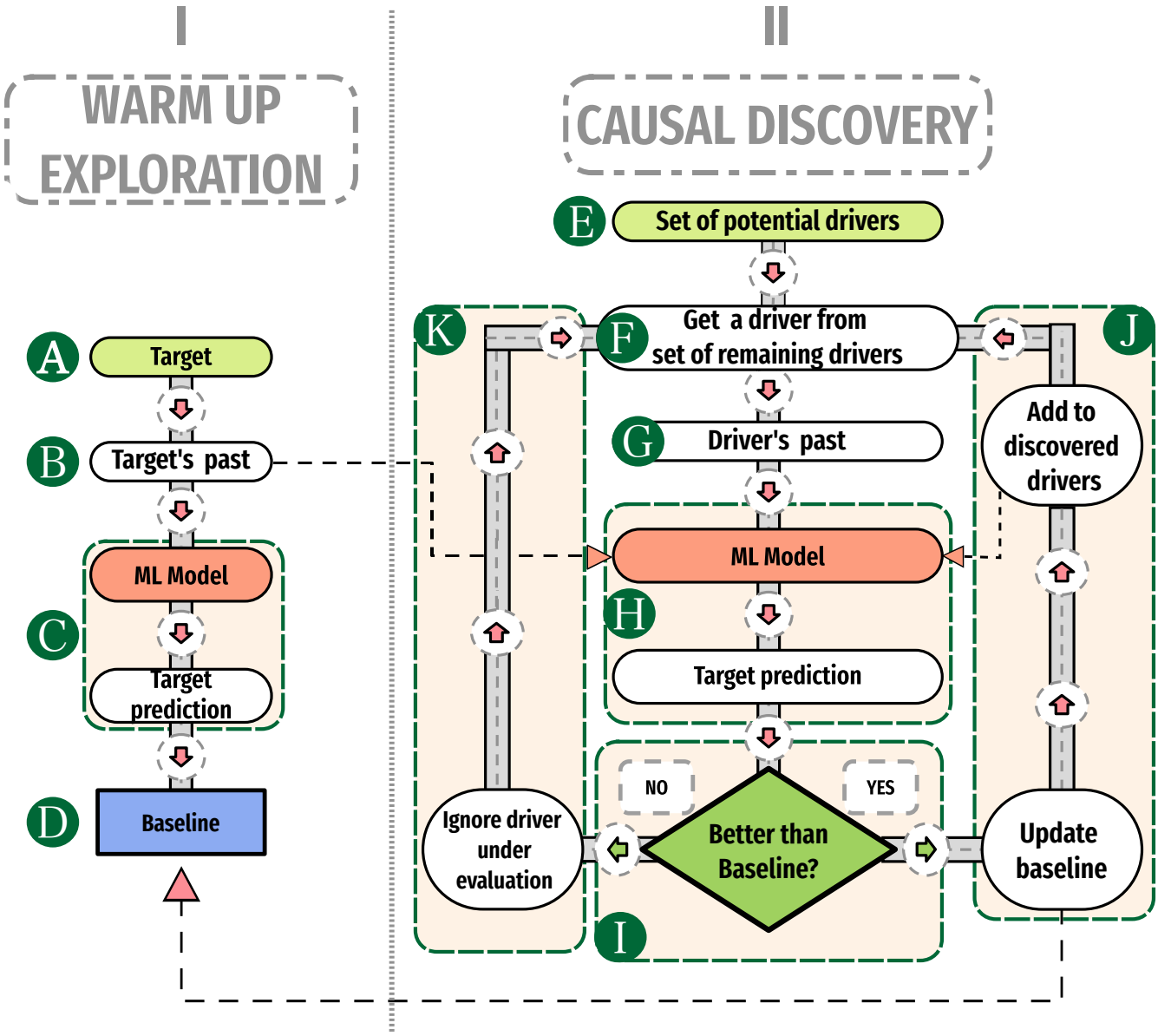

Figure 2. Pipeline used to perform causal discovery. Firstly, a model is built considering only past-lagged values of the target itself, which defines the forecasting baseline (I). The second step is to build a model considering information from both the target and a potential driver that may causally affect the target (II). If there is an improvement in the forecasting, the driver is kept as a predictor (added to the list of discovered drivers, and the baseline is updated). When a new potential driver is tested, information from the already tested drivers that causally affect the target (discovered drivers) is also included.

We train the model and retrieve the feature importance provided by the ensemble model, repeating the process n times and building the set $\mathscr{F}_{\mathbf{X}^d_{t-\tau}} = \{m_1, m_2, \ldots, m_n\}$ per driver d, per feature/lag τ. Similarly, as for assessing driver significance, we fit the data in $\mathscr{F}_{\mathbf{X}^d_{t-\tau}}$ with a statistical distribution and compute the probability of m_{real} of the feature without shuffling being obtained by chance, considering a significance level α.

Combining time series

Depending on the research area we seek to establish causality, time series could be very similar among them. One of the reasons is that the system is constantly controlled (like the case of the oil industry that, for technical reasons, values of certain variables must be kept within a given range). Let us define as an *entity* a driver or target for which we have more than one kind of measurement (information from more than one variable from the same entity). It may occur that time series for the same variable and from different entities are similar, for instance, in terms of Pearson's correlation coefficient. Since we aim to

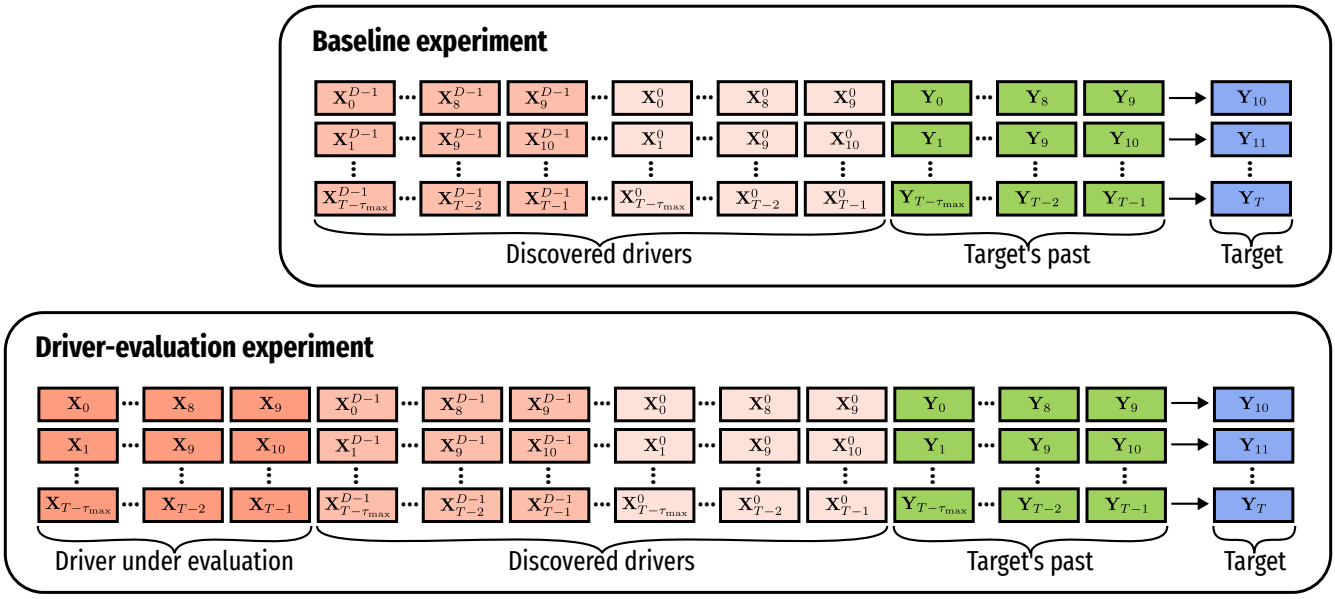

Figure 3. Representation of how feature vectors are defined through a sliding window to assess a causal link from a driver. The first set of features is defined using $\tau_{\max}=10$ past-lagged values — at first from the target $\mathbf{Y}$ only, then also with all the D already discovered drivers $\mathbf{X}^0,\ldots,\mathbf{X}^{D-1}$ — to predict the current target's value $\mathbf{Y}_t$ (**Baseline experiment**). The second set of features is defined considering $\tau_{\max}$ past-lagged values from the target variable $\mathbf{Y}$ as well as from discovered drivers $\mathbf{X}^0,\ldots,\mathbf{X}^{D-1}$ along with values from driver $\mathbf{X}$ under evaluation to predict the current target's value $\mathbf{Y}_t$ (**Driver-evaluation experiment**). When the forecasting of the current target's value $\mathbf{Y}_t$ significantly improves over the baseline by considering data from a new driver $\mathbf{X}$, we consider that driver causally affects the target.

obtain a unique time series representation per entity, we propose combining the time series from the same entity into one in order to circumvent this drawback.

We propose using Uniform Manifold Approximation and Projection (UMAP)[29] to obtain time series representation per entity. In this approach, we perform a random grid-search experimenting with different combinations of the UMAP's hyperparameters *number of neighbors*, *minimum distance* and *spread*. The same setup is applied to every available time series per entity. To determine which UMAP's hyperparameters combination leads to the best results in terms of establishing causal relationships, we used the output of the combination as input in the pipeline defined in Figure 2, find the causal drivers and compare them against a ground truth (if available).

In Figure 4, we show the pipeline followed to combine time series (if needed). Given a dataset with more than one measurement per entity and, if we want to obtain a unique time series representation, we proceed as shown in the lower block. Time series are filtered and those from the same entity are combined into one. The output is then used as a driver or target in Figure 2. In case the dataset has only one measurement per entity, filtering is applied to the time series and is ready to be used as a target/driver.

In this paper, we used this approach to combine time series for oil field related datasets: Time series among various injector wells are correlated to each other because the system is human-controlled. Figure 5 shows two examples of similarity among pairs of injectors' time series, in which similarity is quantified by calculating Pearson's correlation coefficient. The correlation coefficients are high: $\sim$0.8 for the Bottom Hole Pressure (BHP)'s time series and $\sim$0.6 for water injection. In panel 5a is shown both for the time series of BHP and water injection from injectors `IRK004` and `IRK029`; similarly applies to injectors `IRK049` and `IRK036` shown in panel 5b. This high similarity prevents establishing any causal connection between injectors and producers because it is hard to differentiate the injectors' time series from each other.

Step-by-step toy example

To test our algorithms preliminarily (see Section **Proposed methodology**) and, as a proof of concept, we devised a synthetic example to perform time series causal discovery.

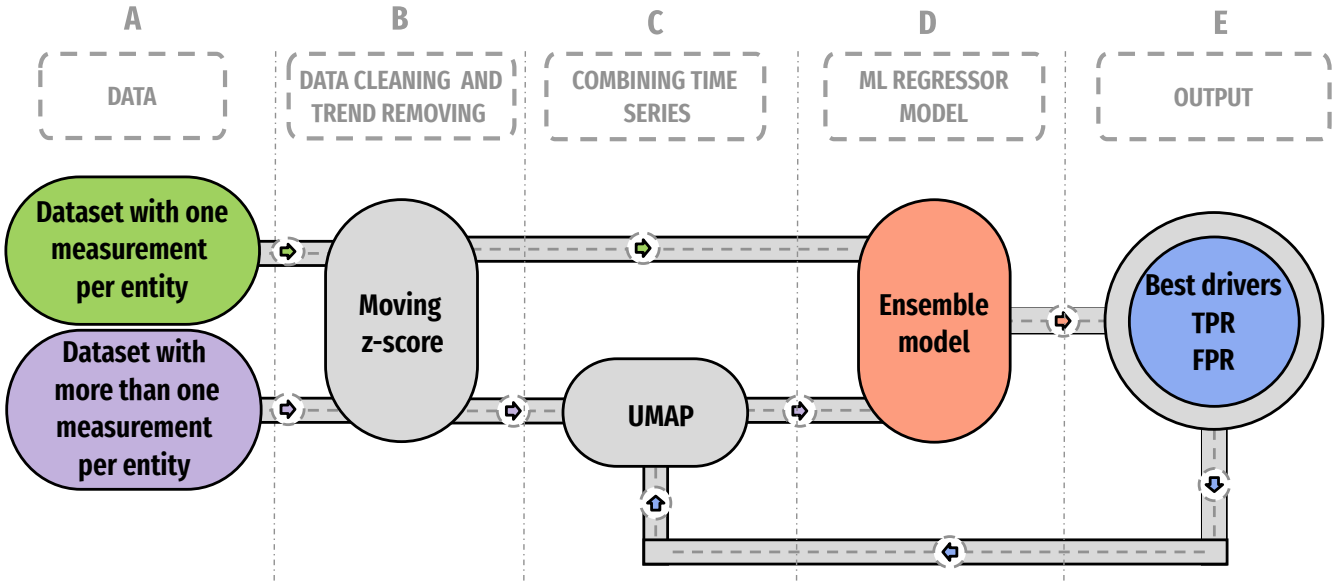

Figure 4. Pipeline for processing and combining (if necessary) time series prior to using them as drivers or targets. (A) Data selection is performed considering the best time windows. (B) Data cleaning and trend removal are applied to the time series on the selected time windows only to keep variability. (C) For the specific case of more than one measurement per entity, time series from the same entity are combined into one using UMAP. (D) Time series from target and drivers are used to assess causality; the pipeline given in Figure 2 fits in this block. (E) The outcome of the ensemble model provides the answer if the driver causally affects the target. If so, the causal driver is compared against the ground truth (if available) to create the True Positive Rate (TPR) and False Positive Rate (FPR) metrics.

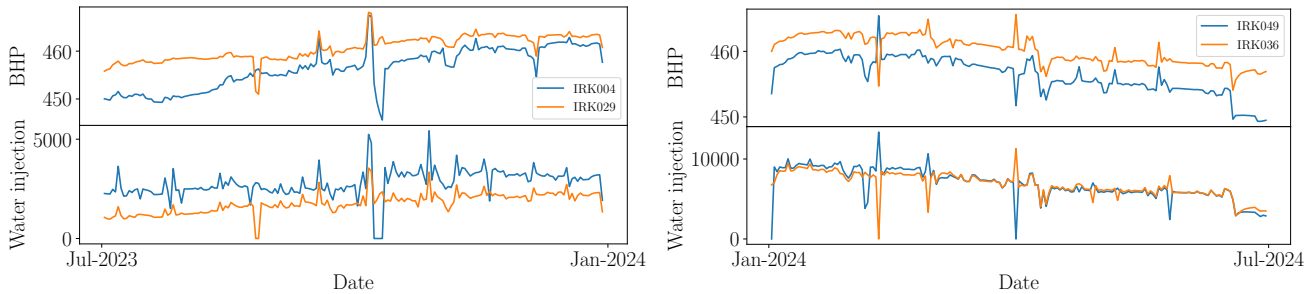

Figure 5. Example of time series similarity. Panels (5a) and (5b) show examples of both BHP and water injection rate, respectively, for two different pairs (one pair per panel) of injector wells in a simulated dataset. In those plots, we can see how similar the water injection time series are for each pair of compared wells. A similar degree of similarity can be seen in the BHP's time series and this similarity is quantified through Pearson's correlation coefficient. Similar behavior is found in time series from a real oil field related dataset.

Example of establishing causal relationships

In equation (2), we show the mathematical definition — following a VAR model — of each time series and which ones are causally affected (namely $\mathbf{X}^1$, $\mathbf{X}^3$ and $\mathbf{X}^7$) by the others. For this example, we consider as targets only $\mathbf{X}^1$ and $\mathbf{X}^7$. The reason for this mathematical construction is to test if we are able to distinguish direct and indirect effects from the drivers: $\mathbf{X}^3$ is affected by $\mathbf{X}^2$, meaning $\mathbf{X}^2$ has an indirect effect (through $\mathbf{X}^3$) on $\mathbf{X}^1$, we want to test if our method can pinpoint $\mathbf{X}^3$ as a causal driver, but not $\mathbf{X}^2$ on $\mathbf{X}^1$. For each time series $\mathbf{X}$, the subindex $t-\tau$ (τ being an integer) means the lag dependence (past values) of the time series' current value $\mathbf{X}_t$. As shown in equation (2), most of the time series depend on themselves, but for those time

series being causally affected by the others, the dependence is with both itself and the other time series.

$$\mathbf{X}^0_t = 0.3\mathbf{X}^0_{t-4} + 0.4\mathbf{X}^0_{t-2} + \eta^0 \tag{2a}$$

$$\mathbf{X}^1_t = 0.4\mathbf{X}^1_{t-4} + 0.5\mathbf{X}^3_{t-4} + 0.3(\mathbf{X}^3_{t-2})^2 + 0.4\mathbf{X}^0_{t-5} + 0.3(\mathbf{X}^0_{t-9})^2 + 0.3(\mathbf{X}^4_{t-1})^2 + 0.4\mathbf{X}^4_{t-7} + 0.4\mathbf{X}^6_{t-3} + 0.6\mathbf{X}^6_{t-8} + \eta^1 \tag{2b}$$

$$\mathbf{X}^2_t = 0.4\mathbf{X}^2_{t-1} + 0.2\mathbf{X}^2_{t-2} + \eta^2 \tag{2c}$$

$$\mathbf{X}^3_t = 0.1\mathbf{X}^3_{t-3} + 0.3\mathbf{X}^3_{t-2} + 0.4\mathbf{X}^2_{t-6} + \eta^3 \tag{2d}$$

$$\mathbf{X}^4_t = 0.3\mathbf{X}^4_{t-8} + 0.3\mathbf{X}^4_{t-1} + \eta^4 \tag{2e}$$

$$\mathbf{X}^5_t = 0.5\mathbf{X}^5_{t-5} + 0.2\mathbf{X}^5_{t-9} + \eta^5 \tag{2f}$$

$$\mathbf{X}^6_t = 0.4\mathbf{X}^6_{t-8} + 0.4\mathbf{X}^6_{t-9} + \eta^6 \tag{2g}$$

$$\mathbf{X}^7_t = 0.5\mathbf{X}^7_{t-9} + 0.8\mathbf{X}^4_{t-1} + 0.7\mathbf{X}^4_{t-6} + 0.4\mathbf{X}^5_{t-7} + 0.6\mathbf{X}^5_{t-8} + 0.9\mathbf{X}^8_{t-1} + 0.7(\mathbf{X}^8_{t-2})^2 + 0.7\mathbf{X}^9_{t-3} + 0.8\mathbf{X}^9_{t-4} + \eta^7 \tag{2h}$$

$$\mathbf{X}^8_t = 0.1\mathbf{X}^8_{t-1} + 0.1\mathbf{X}^8_{t-2} + \eta^8 \tag{2i}$$

$$\mathbf{X}^9_t = 0.3\mathbf{X}^9_{t-7} + 0.5\mathbf{X}^9_{t-8} + \eta^9 \tag{2j}$$

Figure 6 shows the graph with the ground truth of the causal links between the time series, where the label on each edge shows the dependence lags between the target and the driver. Figure 7 presents the time series given by equation (2); for visualization, a short number of data points were generated. This figure does not show seasonality or trend in the time series. Figure 8 shows a few cases of comparison between time series by computing Pearson correlation. Each panel shows the Pearson's cross-correlation coefficient between the time series $\mathbf{X}^1$ or $\mathbf{X}^7$ (columns) with itself or potential drivers (rows) at different lags: There is no strong evidence (looking only at correlations) of causal relationships among the variables, even if we know they exist by definition.

Our algorithm applied to this synthetic example works as described below (mapping the steps in Figure 2):

(A) Given the set of variables $\{\mathbf{X}^0, \mathbf{X}^1, \ldots, \mathbf{X}^9\}$, we select $\mathbf{X}^1$ as potential target (the same procedure applies for $\mathbf{X}^7$ when it is considered as target).

(B) Using past data from $\mathbf{X}^1$, we define feature vectors with past-lagged values $\mathbf{X}^1_{t-\tau}$, $0 < \tau \leq \tau_{\max}$, in which $\tau_{\max}$=10.

(C) An ensemble ML model is trained based on the features $\mathbf{X}^1_{t-\tau}$ and used to predict $\hat{\mathbf{X}}^1_t$ as being the target's current value $\mathbf{X}^1_t$.

(D) An initial *baseline* is established as the evaluation of a metric on the forecasting $\hat{\mathbf{X}}^1_t$ in comparison with the ground-truth $\mathbf{X}^1_t$.

(E) A *set of potential drivers* $\mathscr{D}^p = \{\mathbf{X}^0, \mathbf{X}^2, \ldots, \mathbf{X}^9\}$ that might causally affect $\mathbf{X}^1$ is defined.

(F) If $\mathscr{D}^p$ is not empty, a driver $\mathbf{X}^d$ is selected for consideration and removed from that set, otherwise the result of the pipeline is the set $\mathscr{D}$ of *discovered drivers* incremented through step (J).

(G) Using past data from $\mathbf{X}^d$, we define feature vectors with past-lagged values $\mathbf{X}^d_{t-\tau}$, $0 < \tau \leq \tau_{\max}$.

(H) Features vectors $\mathbf{X}^1_{t-\tau}$, features vectors from already *discovered drivers* in $\mathscr{D}$, and feature vectors $\mathbf{X}^d_{t-\tau}$ are used to predict $\hat{\mathbf{X}}^1_t$.

(I) The metric is evaluated based on $\hat{\mathbf{X}}^1_t$ against $\mathbf{X}^1_t$ and compared with the *baseline*. If the forecasting of the model of step (H) is better than the *baseline*, we go through step (J), otherwise we go through step (K). In either case, we return to step (F).

(J) We state $\mathbf{X}^d$ causally affects $\mathbf{X}^1$, then the *baseline* is updated with the metric evaluated in step (I) and $\mathbf{X}^d$ is added to the set $\mathscr{D}$ of *discovered drivers*.

(K) We consider that $\mathbf{X}^d$ does not causally affect $\mathbf{X}^1$, so we ignore it.

Once every driver in $\mathscr{D}^p$ is tested and, if $\mathscr{D}$ ends up non-empty, the features from the drivers in $\mathscr{D}$ and from the target itself are tested for significance, as is explained in Section **Assessing feature significance**, to determine the most significant features (lags) as shown in Figure 9. In that figure, we can see that the more prominent features match the lags we defined in equation (2) and shown in Figure 6.

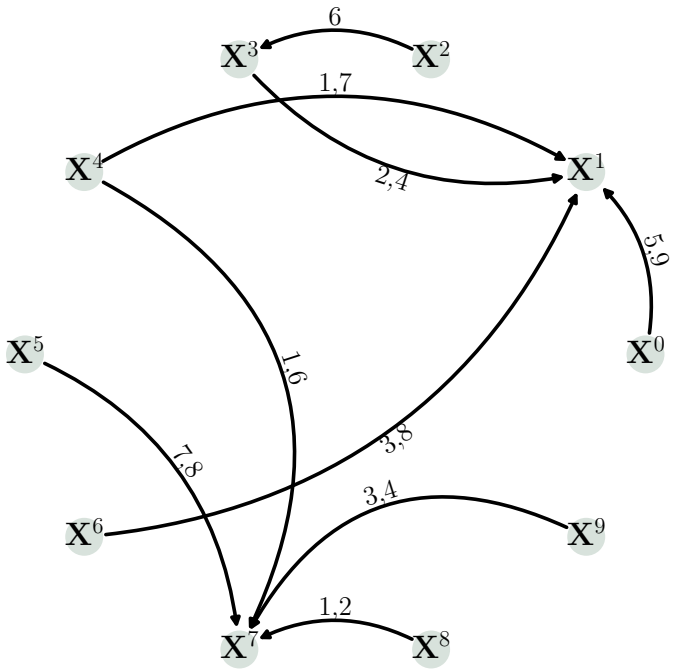

Figure 6. Graph of connections for the synthetic example given by equation (2). The label on the edge between every two nodes means the lag dependence of the target related to the driver.

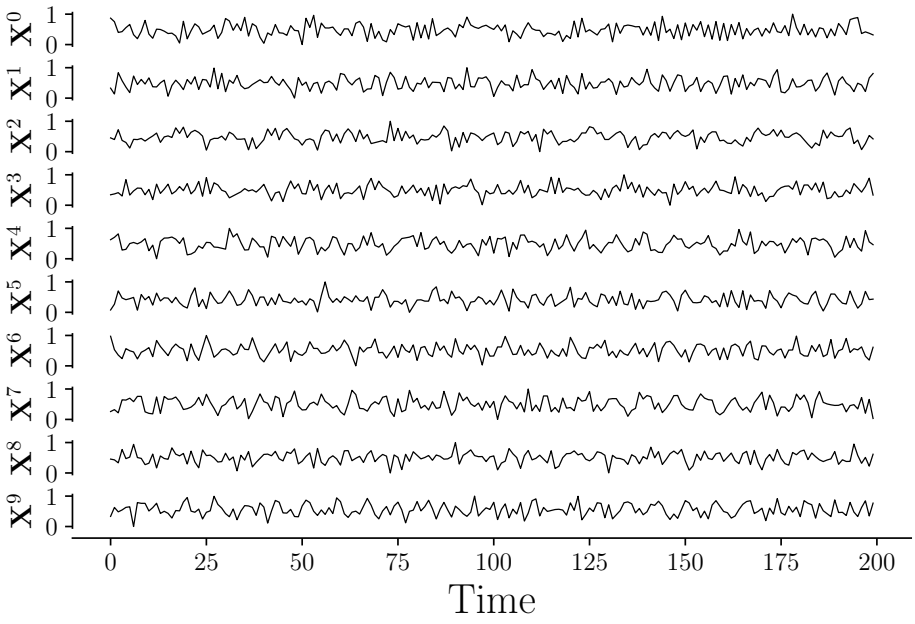

Figure 7. Representation of time series given by equation (2). Each time series was generated considering a VAR model of a different order.

Applications

We tested our algorithms in different datasets to check causal relationships between time series. In subsection **Climate**, we show the results of applying causality tests to weather-related time series. Subsection **Synthetic oil production field** provides the results of applying causality in time series related to a synthetic reservoir. Finally, in subsection **Actual oil production field**, we present results of applying causality to a real production dataset from a Pre-Salt oil field.

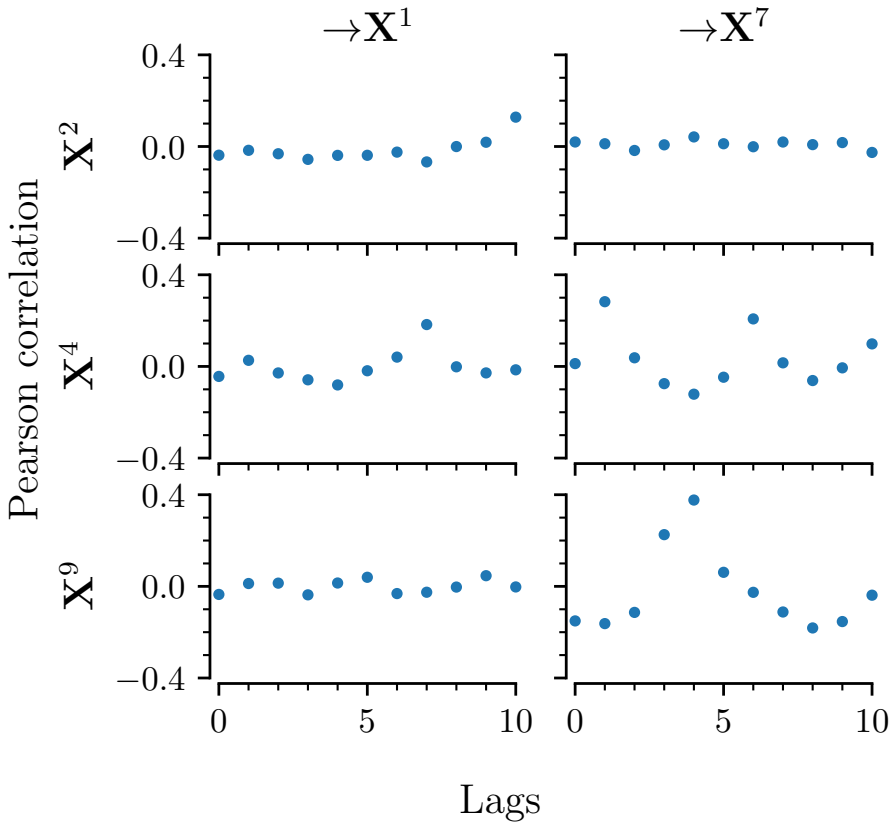

Figure 8. Pearson correlation at different lags for a few of the time series given by equation (2). There is no strong correlation (close to 1) of either $\mathbf{X}^1$ or $\mathbf{X}^7$ with some of their drivers (rows), even though we know by definition in equation (2) there is an influence from the drivers. This shows that, even with a low correlation between drivers and the target, this does not mean the absence of causal relationships.

Climate

To test our causality methods on data from different fields of science, we use the data available by Runge et al.[30]. From the platform CauseMe.net, we chose dataset `TestWEATHnoise` which has four sub-datasets that define the number of variables (N) and length (T, number of observations) in each file (see Section **Data availability**); we employed the sub-dataset with N=5 and T=2000. The description of this dataset provided by the authors says:

> *These weather-type datasets feature typical weather data challenges (autocorrelation, nonlinearity, and time delays) combined with two common computational/statistical challenges, namely high dimensionality and short to large sample sizes. This data is additionally contaminated with observational noise.*

To validate our algorithm in this dataset, the platform requires that the results must be submitted in an already defined format. To our knowledge, the ground truth is not available for download. Our algorithm was applied to each file in the dataset we chose and results were generated for uploading. Once we submit the results, the platform computes the Area Under the ROC Curve (AUC) as a metric to quantify the quality of our method in pinpointing real connections. We obtained an AUC of 0.62 on the dataset `TestWEATHnoise_N-5_T-2000`. Methods proposed by Weichwald et al.[31], like `slarac` (Subsampled Linear Auto-Regression Absolute Coefficients) and `selvar` (Selective auto-regressive model), were tested on the same dataset, obtained AUC of 0.86 and 0.84, respectively (Those values were obtained from `https://causeme.uv.es/rank/` for the same dataset). Those methods are VAR-based and fit linear models on the data, regressing present on past values and inspecting the regression coefficients to decide whether one variable is a Granger-cause of another. Despite the fact that the description of the dataset is stated *nonlinearity* in the data, it seems the dataset has linear relationships across the time series, as supported by the results with `slarac` and `selvar`. Our approach is intended to deal with nonlinear relationships across the time series because our main focus is on working with oil field related data, and one assumption made by the specialists in the oil production area is that the relationship between injector-producer is nonlinear. We obtained good performance in this kind

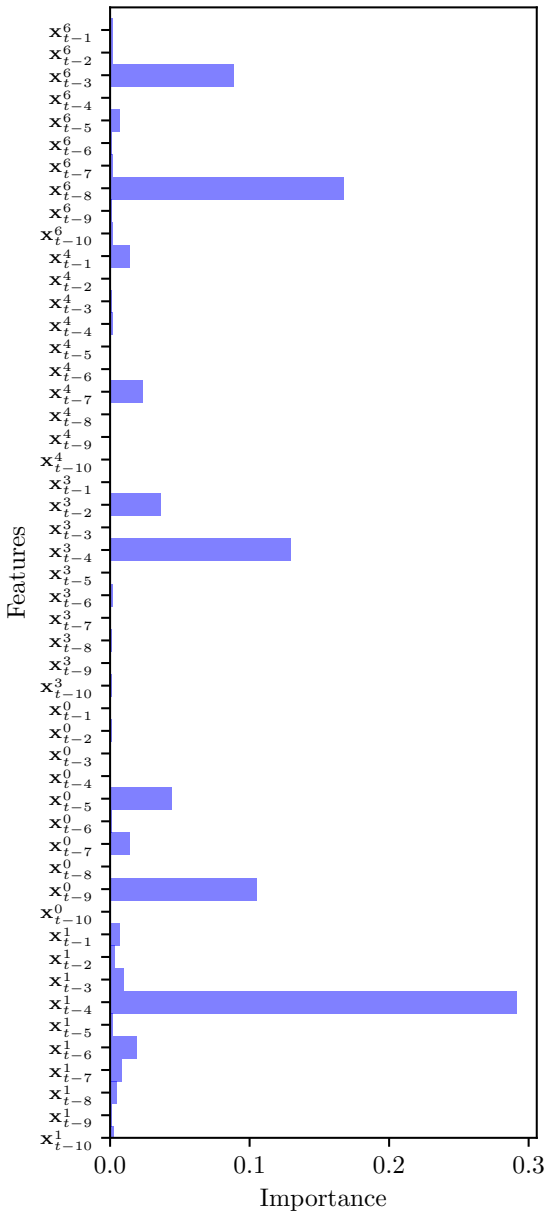

Figure 9. Feature importance for synthetic example with $\mathbf{X}^1$ as target. The features (lags) with higher importance correspond to the drivers' lags influencing the target. For instance, the feature $\mathbf{X}^0_{t-9}$ means the variable $\mathbf{X}^0$ causally affects the variable $\mathbf{X}^1$ nine timestamps later.

of data and this difference in the approaches could lead to different results in the same dataset.

Synthetic oil production field

We tested our algorithms in a synthetic oil field production dataset and 3D models known as UNISIM-II-M-CO (see Section **Data availability**). This dataset represents a typical Pre-Salt field in Brazil[32].

This dataset simulates a reservoir with ten producer wells: `PRK014`, `PRK028`, `PRK045`, `PRK061`, `Wildcat`, `PRK060`, `PRK084`, `PRK085`, and `PRK083`; and eight injector wells: `IRK004`, `IRK028`, `IRK049`, `IRK056`, `IRK063`, `IRK050`, `IRK036`, and `IRK029`, respectively. The first letter on the well name indicates if it is a producer (`P`) or an injector (`I`). The history of oil production is shown in Figure 10, and the history of fluid injection is shown in Figure 11. Those plots are based on the *streamgraph* visualization technique[33] that is meant to show the existence/absence of data over time for each variable; the *y*-axis values cannot be interpreted as real values of material volume (as data were transformed before plotting). As the wells

PRK052 and Wildcat in Figure 10 operated only for short periods, we did not consider them for the analysis resulting in Figure 12. As the injection data of Figure 11 presents Water Alternating Gas (WAG) cycle of 6 months, we chose a six-month time window to perform causality analysis.

For this oil field related time series, we use producers' BHP as targets. The input is both the producers' past BHP and the per-injector combination of the BHP and fluid injection rate (either water or gas, depending on the cycle in which each injector was at the chosen time window) following the pipeline depicted in Figure 4. The results are summarized in Figure 12 and compared with connections confirmed by tracer injection.

In an oil field, tracers (chemical compounds) can be injected into the reservoir to track the underground movement of the fluids[34] between wells. After some time, the chemical tracers which are injected into the injection wells reach some producers and, from this information, injector-producer connections are confirmed. Those connections based on tracer detection are considered our ground truth. Caution must be taken in this regard because the absence of tracer confirmation does not mean that the connection does not exist. There is the possibility that the tracer has not reached the producer yet. This applies to both simulated and real datasets. As seen in Figure 12, there is a good agreement between the connections confirmed by tracer and the ones detected by causal methods. We highlight that our methodology is able to detect connections not yet confirmed by tracer, as the ones to PRK028 from injectors IRK029 and IRK028, respectively, since our analysis depends on pressure communication. Those connections may exist due to the injectors-producer proximity.

The lag dependence between cause and effect, i.e., the time between injection and its impact on the producers, is in the order of days (1–6 days). The travel time we obtained from the tracer data for the simulated dataset shows no clear trend between travel time (difference between the time of detection at the producer and time of injection in the injector) and interwell distance. The complexity of the subsurface geology within the reservoir could lead to preferential pathways that make interwell connections faster regardless of the distance (for example fault pathways, fracture corridors or other high-permeability features/layers in the reservoir formation). This lack of a clear trend in the travel time makes it hard to validate the lags we estimated from the causal analysis.

Actual oil production field

We applied our methods to a real/private dataset from a Pre-Salt field in Brazil. This field consists of 16 production wells and 16 injection wells. The entire field is split into two regions, low and high, based on the wells' location. Figs. 13 and 14 show a visual representation (not the real values) of both oil production rates (upper panel of each plot) and injection history (water injection rates, gas injection rates, and injection BHP) over time. This representation allows us to see when a particular well was open (producing or injecting) or closed. The low region of Figure 13 has fewer producers and injectors than the high region, making it easier to find a time window in which every well operates. In the high region of Figure 14, there are more wells and the dynamics of injection (three bottom variables) frequently change, making it harder to pick up a time window.

To choose the time windows to apply our algorithms, the main criterion is to guarantee that most of the wells (both producers and injectors) per region were active (i.e., open), had available data and, in the case of well closure, that it was for a short period. In the real field data, the daily oil rates are apportioned (measurements are not made individually by well), so we chose to use daily BHP data for producers (since those measurements are made per well) and either water or gas injection rate together with injection BHP for the injectors. Therefore, we did choose the wells P1, P2, P3, P4, P5, P6, I1, I2, I3, I4, and I5 for the low region and the selected time series spans over a period of $\sim$1 year. In Figure 13, we show the chosen time window for analyzing the wells in the low region. For the high region, the chosen wells were P7, P8, P10, P11, P12, I6, I7, I8, I9, I10, and I11. The chosen time window spans over a period of $\sim$6 months (see Figure 14). Because there are more wells than those in the low region, periods of opening/closing (see Figure 14, the upper panel that shows oil production history) are more diverse, as well as the injection cycles for the injector wells. Consequently, it is harder to choose a larger time window with more wells operating and, in the case of injectors, without switching the injection cycle within the time window. This led us to use a $\sim$six-month time window for the high region. No communication between the high and the low regions is expected, according to the reservoir properties, data and spatial distribution of the wells.

A connectivity map of the field is shown in Figure 15. In this real field, there are Oil Production (OP), WAG and Water Injection (WI) wells. The interwell connections confirmed by either water or gas tracers are plotted as well as the connections detected by our causality method. Anyhow, it is worth noticing that not every OP monitors for tracers. Comparing the connections, we see a good agreement between detected and confirmed connections, only one confirmed connection is missed: to well P8. The behavior within an oil reservoir is highly dynamic, changing in time scales from months to years depending on whether new producers or injectors are drilled. It could happen that, within the chosen time frame, connections to P8 are not strong enough to be detected.

It could be misleading to try to quantify the agreement between detected connections by causality and confirmed connections by tracer data through metrics like True Positive Rate (TPR) and False Positive Rate (FPR), or similar metrics. As stated, we do not know every existing connection in the reservoir to use as our ground truth and perform comparison.

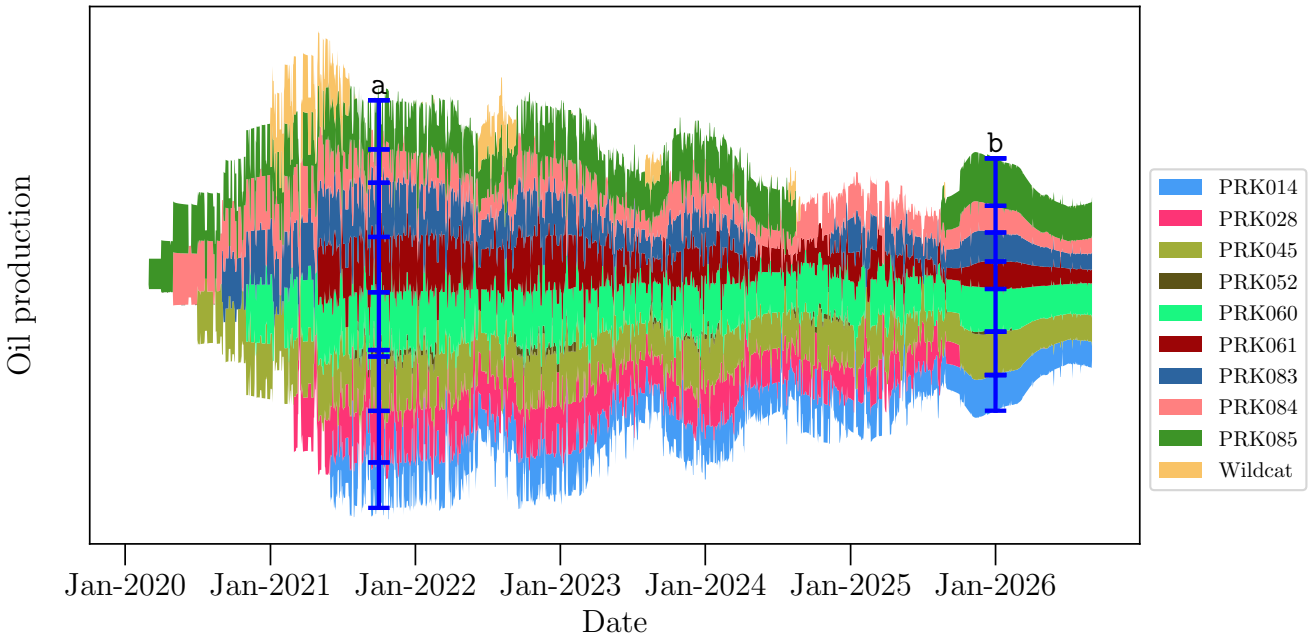

Figure 10. Representation of oil production rate for the UNISIM-II field. Wells `PRK052` and the `Wildcat` operated only for short periods. This kind of data visualization (called *streamgraph*) is meant to show the existence/absence of data over time, each time series represented in the plot was normalized. **How to read the plot**: Given a date (x-axis), the vertical thickness of each time series representation means the relative oil production for each producer compared to the others. In the timestamp labeled as `a`, we can see there were nine producers active and that the producer `PRK052` was the one producing the less; similarly, in the timestamp labeled as `b`, there were seven producers active, the well `PRK028` was closed and `PRK014` was the well with the highest production rate at that time. This methodology applies for each figure in which *streamgraphs* are used to visualize data.

The estimated lags are within a 1–8 day range, as we use pressure-related time series from both injectors and producers, and we expect the communication to be faster by pressure. It is not possible to validate those results due to the lack of sufficient datapoints from tracer information for this real dataset.

Conclusions and future work

This paper showed our approach to establishing causal relationships among time series. The analyses were performed on three datasets: one climate and two oil field related data sets. We performed successful *causal discovery* aiming to estimate which connections exist without quantifying the contribution of each driver in the target it causally affects.

Our approach is data-driven, and we leverage the advantages of ML models that provide the importance of each feature used as input. We use this feature's importance to determine the target's current value-lagged dependence. We also devised a general pipeline to establish causal relationships that can be applied to time series related to different fields.

The proposed solution goes beyond correlations in an attempt to unveil the system's underlying dynamics and the relationships' directionality. Despite the difficulties of having short time windows to work with, we extracted valuable insights regarding interwell connectivity and validated them with the tracer data in the oil field related datasets. Not every interwell connection is confirmed by tracers; therefore, in this scenario, we also depend on the opinion of reservoir engineering specialists to validate our findings. Ideally, it is desirable to have tracer confirmation for every plausible (closed well pairs) connection, however using tracers has an associated economic and operational investment in the applied tracer chemical, equipment, and specialists (involving reservoir engineers, chemists, and field specialists). To our knowledge, this is the first study in the area of oil field production that establishes interwell connectivity based on causal analysis of production data. Understanding and proving the interwell connectivity has a vital impact on reservoir management.

One of the main difficulties we faced was related to the similarity among some time series from different sources (specifically, different injectors). To overcome this, we proposed to combine the time series coming from the same source into one and use this output as a predictor. This solution worked well for the oil field related datasets, providing a way to differentiate time series

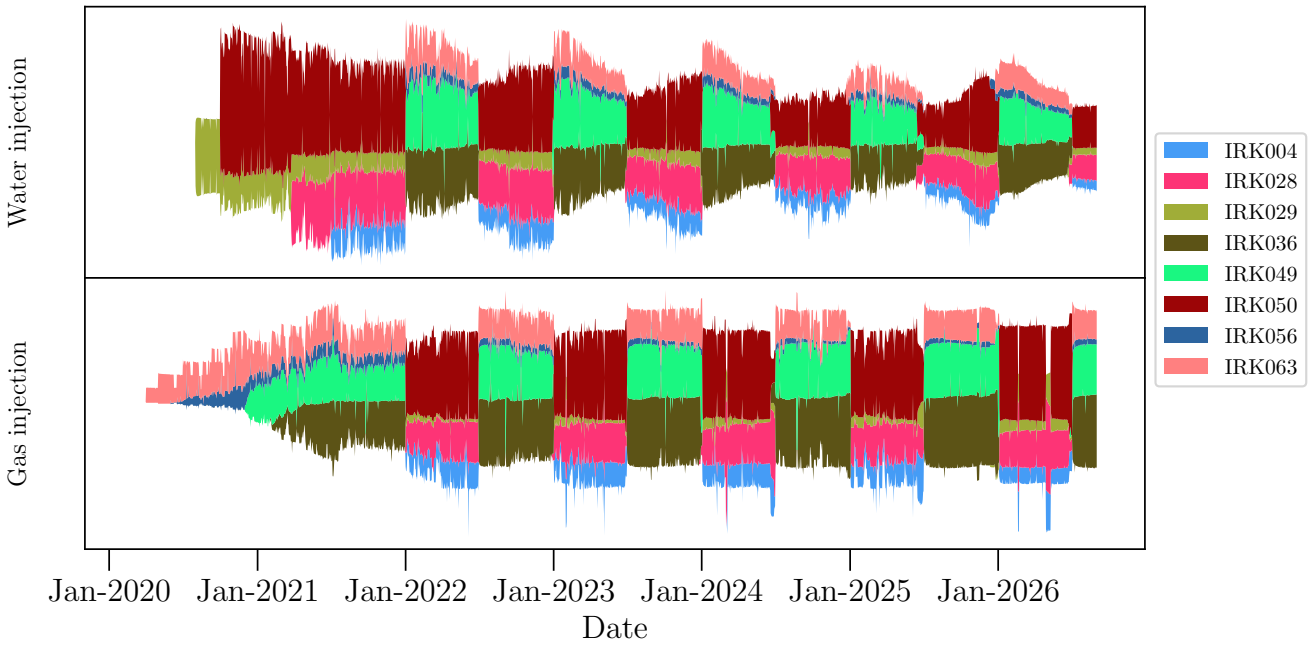

Figure 11. Representation of injection data for the UNISIM-II field/model case. The injection is cyclical, i.e., every six months every injector switches the injected fluid from water to gas or vice-versa.

from different sources.

Future work comprises performing *causal inference*, which allows us to compute the strength of each connection. This is particularly important in the oil field application area because quantifying the strength of the injector-producer connection allows for more fine-grained planning of new strategies of injection or production, depending on the quantification of established connections and gives an independent way to validate connections and use these to improve 3D reservoir models. In this way, we provide an alternative to quantifying the strength of the connection; this quantification is traditionally made using Capacitance-Resistance Model (CRM).

Additionally, another research front is to perform *counterfactuals* analysis, i.e., test what would happen if changes are made to the driver's data: Does the connection to the target hold? Does the strength change? The overall goal is to provide interpretability to some of our ML-based causal algorithms.

Data availability

Both the datasets for climate and UNISIM-II, used in this paper, are available for public access:

Climate dataset: Data downloaded from the platform CauseMe.net. The dataset is available at `https://causeme.uv.es/model/TestWEATHnoise/` (registration required).

UNISIM-II dataset: Available by UNISIM group at the University of Campinas at `https://www.unisim.cepetro.unicamp.br/benchmarks/en/unisim-ii/overview`.

References

1. Pearl, J. Causal inference in statistics: An overview. *Stat. Surv.* **3**, 96–146, DOI: 10.1214/09-SS057 (2009).
2. Runge, J., Nowack, P., Kretschmer, M., Flaxman, S. & Sejdinovic, D. Detecting and quantifying causal associations in large nonlinear time series datasets. *Sci. Adv.* **5**, eaau4996, DOI: 10.1126/sciadv.aau4996 (2019).
3. Pearl, J. & Mackenzie, D. *The Book of Why: The New Science of Cause and Effect* (Basic Books, Inc., New York, NY, USA, 2018), 1st edn.
4. Runge, J. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdiscip. J. Nonlinear Sci.* **28**, 075310, DOI: 10.1063/1.5025050 (2018).
5. Nauta, M., Bucur, D. & Seifert, C. Causal discovery with attention-based convolutional neural networks. *Mach. Learn. Knowl. Extr.* **1**, 312–340, DOI: 10.3390/make1010019 (2019).

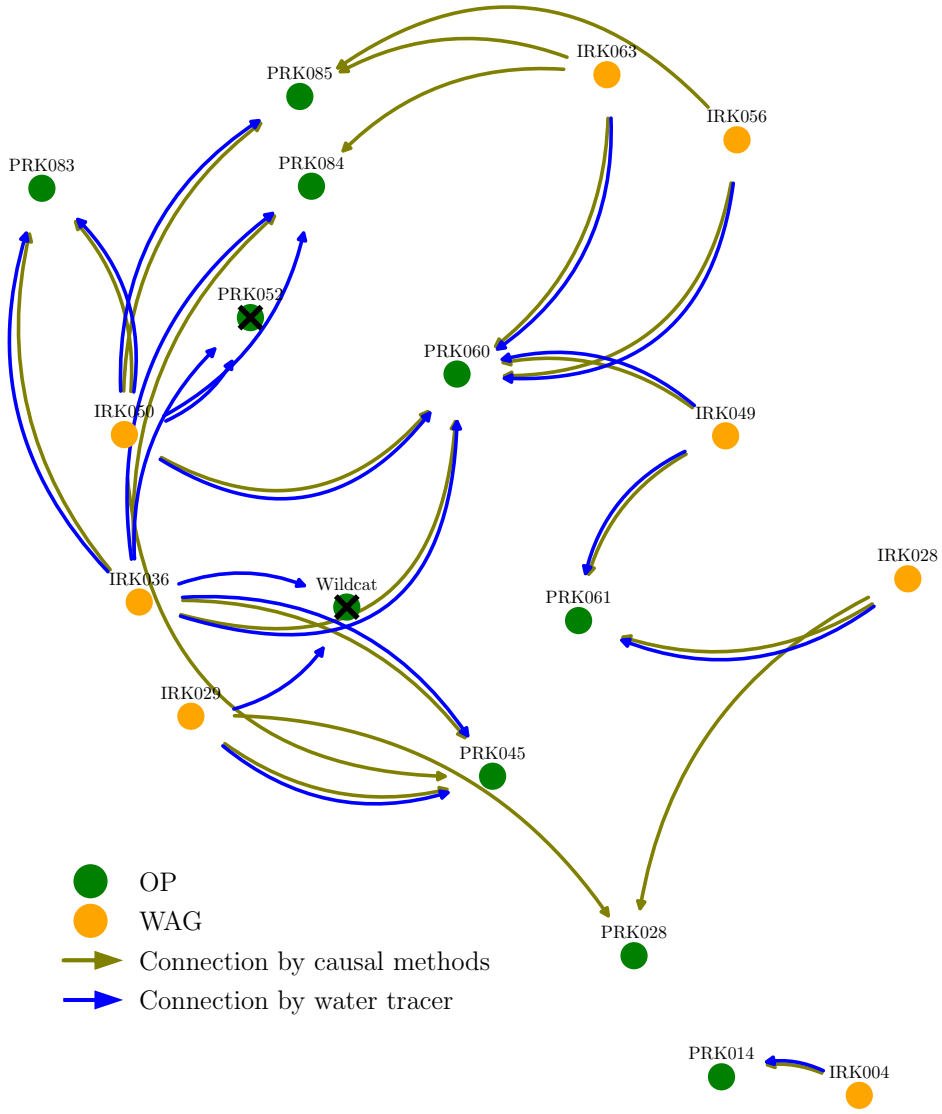

Figure 12. Connectivity map based on causality analysis applied to the UNISIM-II dataset. Connections detected by causality analysis are shown (olive arrows) and compared against connections confirmed by water tracer (blue arrows). The cross mark represents wells not included in the analysis.

6. Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J. & Schölkopf, B. Distinguishing cause from effect using observational data: Methods and benchmarks. *J. Mach. Learn. Res.* **17**, 1–102 (2016).
7. Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* **62**, 54–60, DOI: 10.1145/3241036 (2019).
8. Lines, J. & Bagnall, A. Time series classification with ensembles of elastic distance measures. *Data Min. Knowl. Discov.* **29**, 565–592 (2015).
9. Ma, Q., Zheng, J., Li, S. & Cottrell, G. W. Learning representations for time series clustering. In *Advances in Neural Information Processing Systems*, vol. 32, 3781–3791 (Curran Associates, Inc., 2019).
10. Wang, Y. *et al.* Deep factors for forecasting. In *International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, 6607–6617 (2019).
11. de Oliveira Werneck, R. *et al.* Data-driven deep-learning forecasting for oil production and pressure. *J. Petroleum Sci. Eng.* **210**, 109937, DOI: https://doi.org/10.1016/j.petrol.2021.109937 (2022).
12. Moraffah, R. *et al.* Causal inference for time series analysis: problems, methods and evaluation. *Knowl. Inf. Syst.* **63**, 3041–3085, DOI: 10.1007/s10115-021-01621-0 (2021).

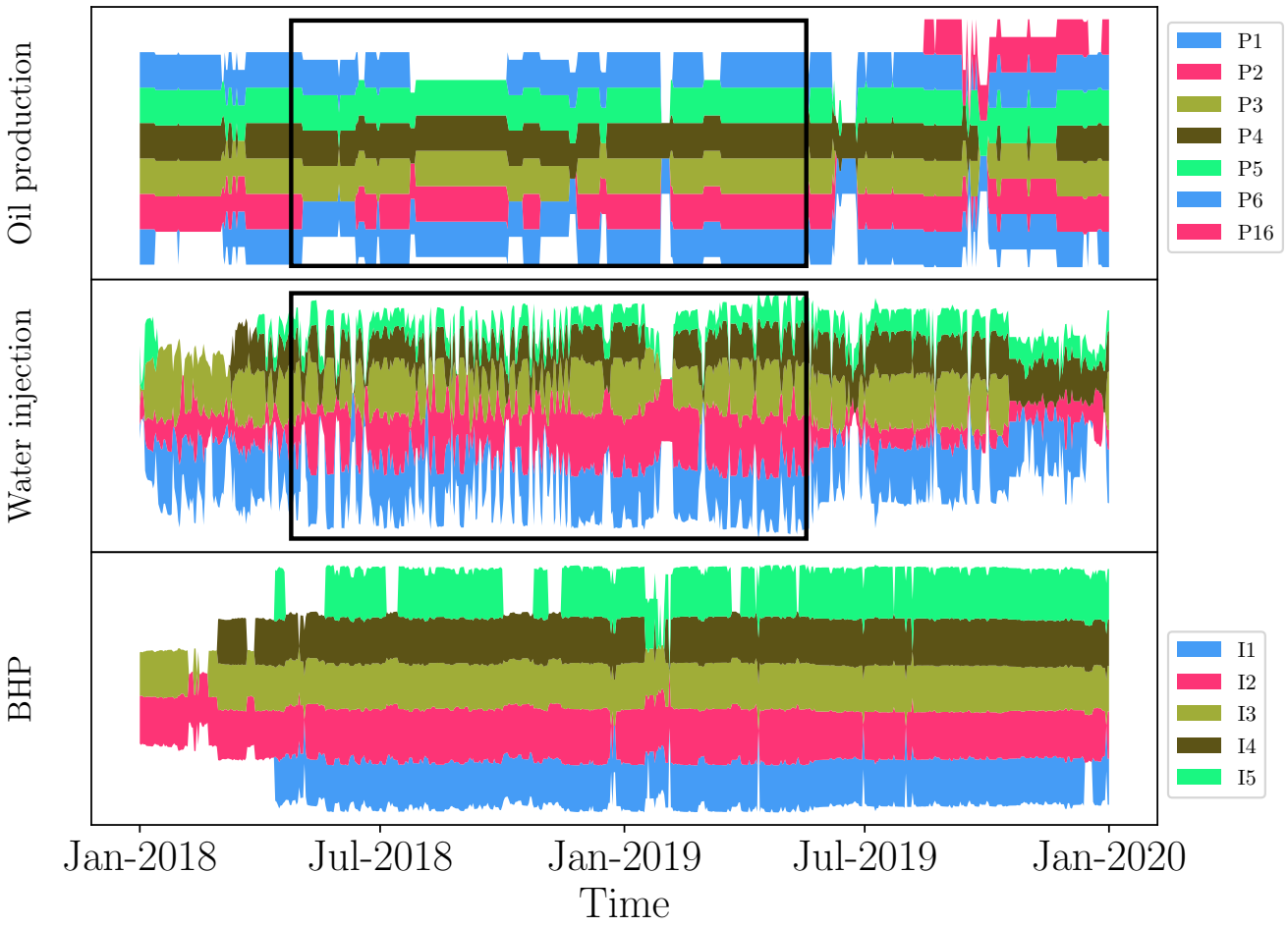

Figure 13. Visualization of available data for the low region of the Pre-Salt oil production field. The first panel represents data coming from the producers and the remaining panels represent data coming from injectors.

13. Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438 (1969).

14. Haufe, S., Müller, K.-R., Nolte, G. & Krämer, N. Sparse causal discovery in multivariate time series. In *Workshop on Causality: Objectives and Assessment at NIPS 2008*, vol. 6 of *Proceedings of Machine Learning Research*, 97–106 (PMLR, Whistler, Canada, 2010).

15. Siggiridou, E. & Kugiumtzis, D. Granger causality in multivariate time series using a time-ordered restricted vector autoregressive model. *IEEE Transactions on Signal Process.* **64**, 1759–1773, DOI: 10.1109/TSP.2015.2500893 (2016).

16. Spirtes, P., Glymour, C. & Scheines, R. *Causation, Prediction, and Search* (MIT press, 2000), 2nd edn.

17. Chu, T. & Glymour, C. Search for additive nonlinear time series causal models. *J. Mach. Learn. Res.* **9**, 967–991 (2008).

18. Runge, J. Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets. In *Conference on Uncertainty in Artificial Intelligence*, vol. 124 of *Proceedings of Machine Learning Research*, 1388–1397 (2020).

19. Ullman, J. B. & Bentler, P. M. Structural equation modeling. In *Handbook of Psychology*, chap. 23, DOI: 10.1002/9781118133880.hop202023 (John Wiley & Sons, Ltd, 2012), 2nd edn.

20. Spirtes, P. & Zhang, K. Causal discovery and inference: concepts and recent methodological advances. *Appl. Informatics* **3**, 3, DOI: 10.1186/s40535-016-0018-x (2016).

21. Peters, J., Janzing, D. & Schölkopf, B. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, vol. 26 (Curran Associates, Inc., 2013).

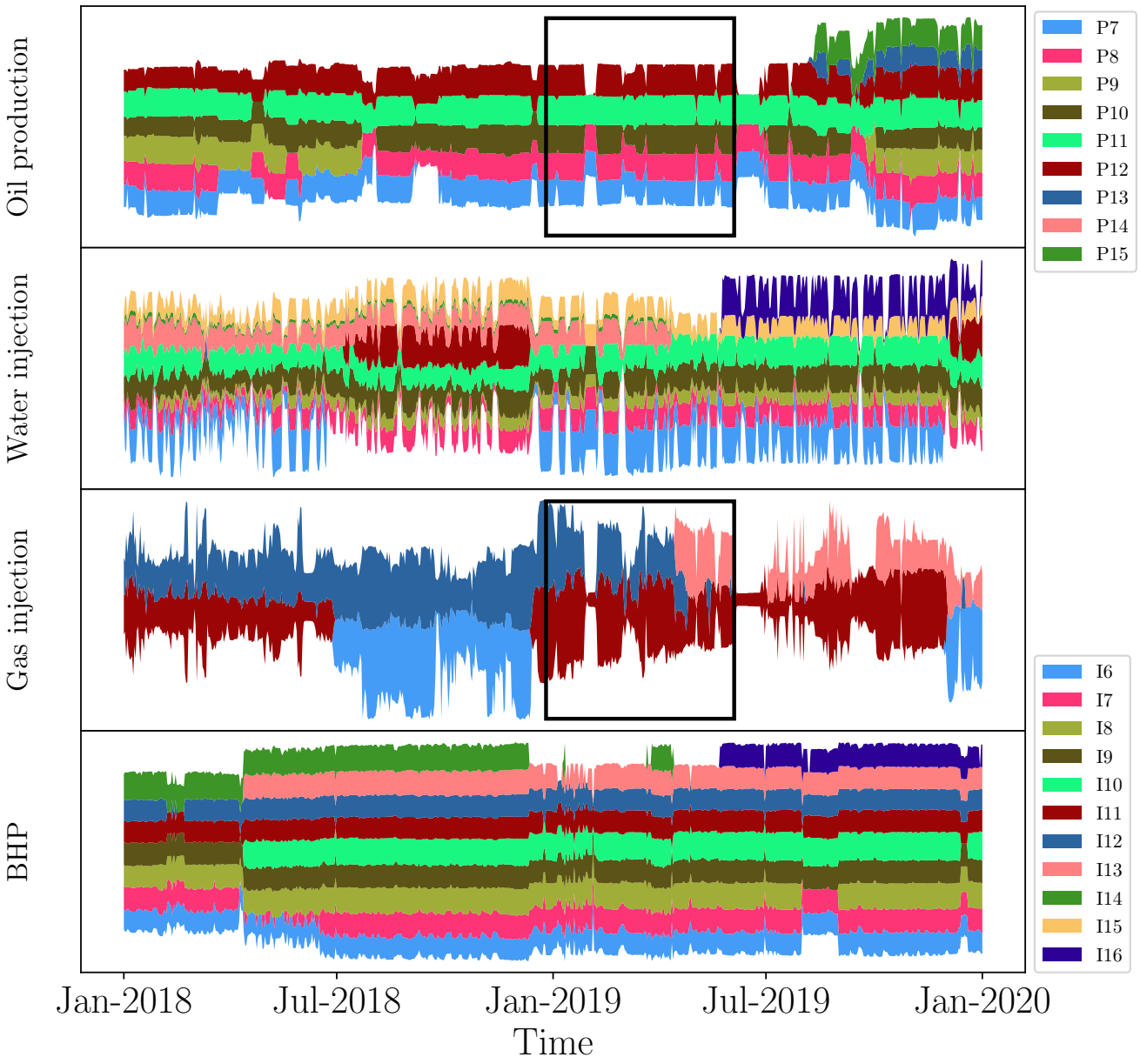

Figure 14. Visualization of available data for the high region of the Pre-Salt oil production field. The first panel represents data from producers, the remaining panels represent data coming from injector including injection BHP.

22. Tank, A., Covert, I., Foti, N., Shojaie, A. & Fox, E. B. Neural granger causality. *IEEE Transactions on Pattern Analysis Mach. Intell.* **44**, 4267–4279, DOI: 10.1109/TPAMI.2021.3065601 (2021).

23. Barić, D., Fumić, P., Horvatić, D. & Lipic, T. Benchmarking attention-based interpretability of deep learning in multivariate time series predictions. *Entropy* **23**, 143, DOI: 10.3390/e23020143 (2021).

24. Mariea, J. P. *Borges, his Aleph, and The Aleph: Constructing identity through the written text* (State University of New York at Buffalo, 2008).

25. Leng, S., Xu, Z. & Ma, H. Reconstructing directional causal networks with random forest: Causality meeting machine learning. *Chaos: An Interdiscip. J. Nonlinear Sci.* **29**, 093130, DOI: 10.1063/1.5120778 (2019).

26. Li, L. *et al.* A causal inference model based on Random Forests to identify the effect of soil moisture on precipitation. *J. Hydrometeorol.* **21**, 1115–1131, DOI: 10.1175/JHM-D-19-0209.1 (2020).

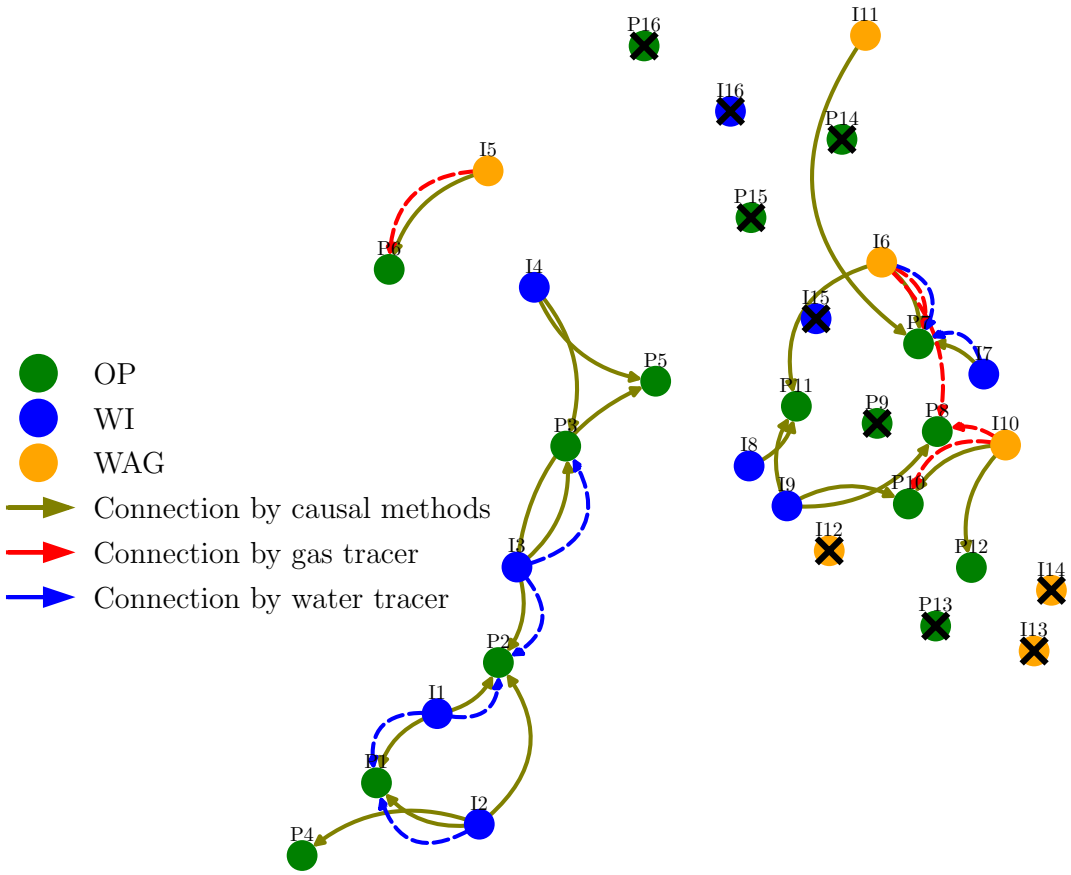

Figure 15. Connectivity map based on causality analysis applied to the Pre-Salt oil production field. The location of each well is shown and so are established connections based on our method and on tracer data. The cross mark represents wells not included in the analysis.

27. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).

28. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost: Unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 6639–6649 (Curran Associates Inc., Red Hook, NY, USA, 2018).

29. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction (2020). 1802.03426.

30. Runge, J. *et al.* Inferring causation from time series in earth system sciences. *Nat. Commun.* **10**, DOI: 10.1038/s41467-019-10105-3 (2019).

31. Weichwald, S. *et al.* Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In Escalante, H. J. & Hadsell, R. (eds.) *Proceedings of the NeurIPS 2019 Competition and Demonstration Track*, vol. 123 of *Proceedings of Machine Learning Research*, 27–36 (PMLR, 2020).

32. Correia, M., Hohendorff, J., Gaspar, A. T. F. S. & Schiozer, D. UNISIM-II-D: Benchmark case proposal based on a carbonate reservoir. In *SPE Latin American and Caribbean Petroleum Engineering Conference*, DOI: 10.2118/177140-ms (Quito, Ecuador, 2015).

33. Byron, L. & Wattenberg, M. Stacked graphs – geometry & aesthetics. *IEEE Transactions on Vis. Comput. Graph.* **14**, 1245–1252, DOI: 10.1109/TVCG.2008.166 (2008).

34. Du, Y. & Guan, L. Interwell tracer tests: Lessons learnted from past field studies. In *SPE Asia Pacific Oil and Gas Conference and Exhibition*, 93140, DOI: 10.2118/93140-ms (Jakarta, Indonesia, 2005).

Acknowledgments

This research was carried out as part of the ongoing R&D project with Shell Brasil Petróleo Ltda, registered as ANP number 21373-6 as “Desenvolvimento de Técnicas de Aprendizado de Máquina para Análise de Dados Complexos de Produção de um Campo do Pré-Sal” — ”Machine-Learning Development for Analysis of Complex Production Data in a Pre-Salt Carbonate Field” — (UNICAMP/Shell Brazil/ANP) funded by Shell Brazil Technology, under the ANP R&D levy as “Compromisso de Investimentos com Pesquisa e Desenvolvimento”. We thank Frances Abbots for the valuable review of this work and Shell Brazil for permission to publish this work.

Author contributions

M.C. conceived and performed the experiments, and led the paper’s writing. A.F. provided insights on the execution of the experiments, analysis of the results and assisted with the paper’s writing. P.R.M.J. helped with the paper’s writing and designed some diagrams. A.S.V. provided insights on the diagrams’ design. All the authors contributed to revising the manuscript, providing comments and/or suggestions to improve it and approving its final version.

Competing interests

The authors declare no competing interests.