Causal deep learning models for studying the Earth system

Tobias Tesch[1,2], Stefan Kollet[1,2], and Jochen Garcke[3,4]

[1]Institute of Bio- and Geosciences, Agrosphere (IBG-3), Forschungszentrum Jülich, 52425 Jülich, Germany
[2]Center for High-Performance Scientific Computing in Terrestrial Systems, Geoverbund ABC/J, Jülich, Germany
[3]Fraunhofer Center for Machine Learning and Fraunhofer SCAI, 53757 Sankt Augustin, Germany
[4]Institut für Numerische Simulation, Universität Bonn, 53115 Bonn, Germany

Correspondence: Tobias Tesch (t.tesch@fz-juelich.de)

Abstract. Earth is a complex non-linear dynamical system. Despite decades of research, and considerable scientific and methodological progress, many processes and relations between Earth system variables remain poorly understood. Current approaches for studying relations in the Earth system rely either on numerical simulations or statistical approaches. However, there are several inherent limitations to existing approaches, including high computational costs, uncertainties in numerical models, strong assumptions about linearity or locality, and the fallacy of correlation and causality. Here, we propose a novel methodology combining deep learning (DL) and principles of causality research in an attempt to overcome these limitations. On the one hand, we employ the recent idea of training and analyzing DL models to gain new scientific insights into relations between input and target variables. On the other hand, we use that a statistical model learns the causal effect of an input variable on a target variable if suitable additional input variables are included. As an illustrative example, we apply the methodology to study soil moisture-precipitation coupling in ERA5 climate reanalysis data across Europe. We demonstrate that, harnessing the great power and flexibility of DL models, the proposed methodology may yield new scientific insights into complex nonlinear and non-local coupling mechanisms in the Earth system.

1 Introduction

The Earth system comprises many complex processes and non-linear relations between variables that are still not fully understood. Considering for example soil moisture-precipitation coupling, i.e. the question how precipitation changes if soil moisture is changed, it is well-known that soil moisture affects the temperature and humidity profile of the atmosphere and thereby influences the development and onset of precipitation (Seneviratne et al., 2010; Santanello et al., 2018). However, because there are several concurring pathways of soil moisture-precipitation coupling, it remains an open question whether an increase in soil moisture leads to an increase or decrease in precipitation. Answering this question might lead to improved precipitation predictions with numerical models.

Approaches for studying relations in the Earth system may be broadly divided into approaches based on numerical simulations (e.g. Koster, 2004; Seneviratne et al., 2006; Hartick et al., 2021), and statistical approaches (e.g. Taylor, 2015; Guillod et al., 2015; Tuttle and Salvucci, 2016). Both classes of approaches have several inherent limitations. Approaches based on numerical simulations usually have high computational costs and, even more importantly, rely on the correct representation of

the considered relations in the numerical model. For example, precipitation in numerical models lacks accuracy due to several simplified parameterizations, thus, using these models to study soil moisture-precipitation coupling is problematic. On the other hand, statistical approaches usually have much lower computational costs and can directly be applied to observational data. However, current statistical approaches have strong limitations on their own, for example due to assumptions on linearity or locality of considered relations and negligence of the difference between causality and correlation.

A recent statistical approach for studying relations in the Earth system is to train deep learning (DL) models to predict one Earth system variable given one or several others, and use methods from the realm of interpretable DL (Zhang and Zhu, 2018; Montavon et al., 2018; Gilpin et al., 2018; Molnar, 2019; Samek et al., 2021) to analyze the relations learned by the models (Roscher et al., 2020). The approach has been applied in several recent studies (Ham et al., 2019; Gagne II et al., 2019; McGovern et al., 2019; Toms et al., 2020; Ebert-Uphoff and Hilburn, 2020; Padarian et al., 2020), and the use of DL models allows to overcome common assumptions in other statistical approaches like linearity or locality. So far, however, the difference between causality and correlation has been neglected in studies using this approach. Indeed, DL models might learn various (spurious) correlations between input and target variables, while researchers striving for new scientific insights are most interested in causal relations.

Therefore, in this work, we propose extending the approach by combining it with a result from causality research stating that a statistical model may learn the causal effect of an input variable on a target variable if suitable additional input variables are included (Pearl, 2009; Shpitser et al., 2010). In the geosciences, this result has only recently received attention in the work of (Massmann et al., 2021). In this work, it is combined with the methodology of training and analyzing DL models to gain new scientific insights for the first time. Note that there are several other recent studies on causal inference methods in the geosciences (e.g. Tuttle and Salvucci, 2016, 2017; Ebert-Uphoff and Deng, 2017; Green et al., 2017; Runge, 2018; Runge et al., 2019; Barnes et al., 2019; Massmann et al., 2021). However, most of them focus on discovering causal dependencies between variables, while the proposed methodology assumes prior knowledge on causal dependencies and focuses on quantifying the strength and sign of a particular causal dependency. As an illustrative example, we apply the proposed methodology to study soil moisture-precipitation coupling in ERA5 climate reanalysis data across Europe. Other geoscientific questions that could be addressed with the proposed methodology are, for example, soil moisture-temperature coupling (Seneviratne et al., 2006; Schwingshackl et al., 2017; Schumacher et al., 2019) and soil moisture-atmospheric carbon dioxide coupling (Green et al., 2019; Humphrey et al., 2021).

This manuscript is structured as follows: Sect. 2 introduces the background on causality research and details the proposed methodology. Sect. 3 presents the application to soil moisture-precipitation coupling and provides a comparison to other approaches. Finally, Sect. 4 contains several additional analyses to assess the statistical significance and correctness of results obtained with the proposed methodology.

2 Methodology

To introduce the proposed methodology, which combines deep learning with a result from causality research, we first give a basic introduction into the required concepts from causality research. Based on that, we describe how one can train a DL model that reflects causality.

2.1 Background on causality

If we could change the value of any Earth system variable, e.g. increase soil moisture in some area, this would potentially affect numerous other Earth system variables, e.g. evaporation, temperature and precipitation. The variable that was changed thus has a *causal* impact on the latter variables. Formally, the causal effect of some variable $\boldsymbol{X} \in \mathbb{R}^d$ on another variable $\boldsymbol{Y} \in \mathbb{R}^n$ is the expected response of $\boldsymbol{Y}$ to changing the value of $\boldsymbol{X}$. To determine this impact, one has to determine the expected value of $\boldsymbol{Y}$ given that one sets $\boldsymbol{X}$ to some arbitrary value $\boldsymbol{x}$. In the framework of Structural Causal Models (SCMs) introduced below, setting $\boldsymbol{X}$ to $\boldsymbol{x}$ is represented by a mathematical *intervention* operator $do(\boldsymbol{X} = \boldsymbol{x})$, and the sought value is referred to as the *post-intervention* expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})]$.

In some cases, $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})]$ can be determined experimentally by setting $\boldsymbol{X}$ to $\boldsymbol{x}$ while monitoring $\boldsymbol{Y}$. For example, in Earth System Modeling (ESM), one may be able to set $\boldsymbol{X}$ to $\boldsymbol{x}$ in numerical experiments. However, often it is impossible to determine $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})]$ experimentally due to computational constraints or because of the lack of appropriate numerical models. Obviously, analog experiments are even harder to perform or impossible in case of large scale interactions in the Earth system.

The framework of SCMs (Pearl, 2009) provides a deeper understanding of the notion $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})]$, and describes how it can be determined without experimentally setting $\boldsymbol{X}$ to $\boldsymbol{x}$. The framework is briefly introduced in the following. For a more in-depth introduction we refer to (Pearl, 2009). An introduction to the framework in the context of geosciences is given in (Massmann et al., 2021).

2.1.1 Structural Causal Models

In the framework of SCMs, the considered system, e.g. the Earth system, is described by a causal graph and associated structural equations. A causal graph is a directed acyclic graph, in which nodes represent the variables of the system and edges encode the dependencies between these variables. For example, in the system described by Fig. 1a, variable $\boldsymbol{Y}$ depends on all other variables, although the lack of an edge from $\boldsymbol{X}$ to $\boldsymbol{Y}$ implies that $\boldsymbol{X}$ only affects $\boldsymbol{Y}$ indirectly via its impact on $\boldsymbol{C_2}$. Parents of a considered variable (node) are all variables that have a direct effect on that variable, i.e. all variables with an edge pointing to that variable. In the following the terms node and variable are used interchangeably.

Formally, a variable in the causal graph is determined by a function f, whose inputs are its parents and a random variable $\boldsymbol{U}$ representing potential chaos and variables not included in the causal graph explicitly. For example, for the system in Fig. 1a, the four variables are determined by four functions $f_{C_1}, f_{C_2}, f_X, f_Y$:

$$\boldsymbol{C_1} = f_{C_1}(\boldsymbol{U_{C_1}}),\ \ \boldsymbol{X} = f_X(\boldsymbol{C_1}, \boldsymbol{U_X}),\ \ \boldsymbol{C_2} = f_{C_2}(\boldsymbol{X}, \boldsymbol{U_{C_2}}),\ \ \boldsymbol{Y} = f_Y(\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{U_Y}). \tag{1}$$

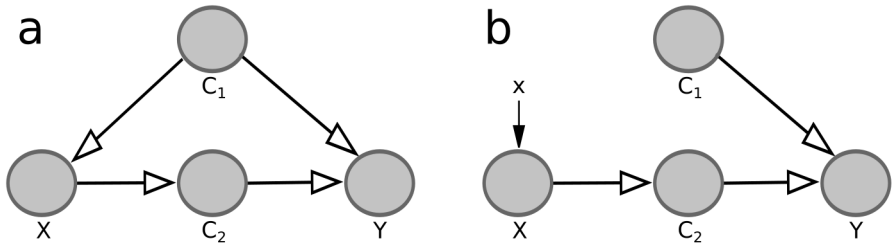

Figure 1. Example for a causal graph (a) and corresponding causal graph for setting variable X to some arbitrary value x (b). The grey circles are referred to as nodes of the graph, while the arrows are referred to as directed edges.

These equations are called structural equations. The random variables $\boldsymbol{U_{C_1}}, \boldsymbol{U_{C_2}}, \boldsymbol{U_X}, \boldsymbol{U_Y}$ are assumed to be mutually independent and give rise to a probability distribution $\mathbb{P}(\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{X}, \boldsymbol{Y})$, which describes the probability of observing any tuple of values $(\boldsymbol{c_1}, \boldsymbol{c_2}, \boldsymbol{x}, \boldsymbol{y})$. Integrating the product of $\boldsymbol{Y}$ and this probability distribution over all tuples $(\boldsymbol{c_1}, \boldsymbol{c_2}, \boldsymbol{y})$ for some fixed value $\boldsymbol{x}$, one obtains the expected value of $\boldsymbol{Y}$ given that one *observes* the value $\boldsymbol{x}$ of $\boldsymbol{X}$, i.e.

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}] = \int\limits_{\boldsymbol{c_1},\boldsymbol{c_2},\boldsymbol{y}} \boldsymbol{y} \cdot \mathbb{P}[\boldsymbol{C_1}=\boldsymbol{c_1}, \boldsymbol{C_2}=\boldsymbol{c_2}, \boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}=\boldsymbol{x}]. \tag{2}$$

As stated above, to determine the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$, one has to determine the expected value of $\boldsymbol{Y}$ given that one *set* $\boldsymbol{X}$ to some arbitrary value $\boldsymbol{x}$, i.e. the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})]$. By setting $\boldsymbol{X}$ to some arbitrary value $\boldsymbol{x}$, all dependencies of $\boldsymbol{X}$ on other variables are eliminated. Within the framework of SCMs, this corresponds to removing all edges in the causal graph pointing to $\boldsymbol{X}$, and modifying the structural equation for $\boldsymbol{X}$ accordingly. For example, when studying the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ in Fig. 1a, the modified system is described by the causal graph in Fig. 1b with the associated structural equations

$$\boldsymbol{C_1} = f_{C_1}(\boldsymbol{U_{C_1}}), \;\; \boldsymbol{X}=\boldsymbol{x}, \;\; \boldsymbol{C_2} = f_{C_2}(\boldsymbol{X}, \boldsymbol{U_{C_2}}), \;\; \boldsymbol{Y} = f_Y(\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{U_Y}). \tag{3}$$

Again, the random variables $\boldsymbol{U_{C_1}}, \boldsymbol{U_{C_2}}, \boldsymbol{U_Y}$ give rise to a probability distribution $\mathbb{P}(\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}))$, referred to as post-intervention probability distribution, and the corresponding post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})]$. This expected value is used to determine the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ and differs from the expected value for the original system, $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}]$. For instance, in the example from Fig. 1, knowing $\boldsymbol{X}$ allows to draw conclusions about $\boldsymbol{Y}$ both in the original system (Fig. 1a) as well as in the modified system (Fig. 1b), because $\boldsymbol{X}$ has a causal effect on $\boldsymbol{Y}$ (via its impact on $\boldsymbol{C_2}$). However, in the original system, knowing $\boldsymbol{X}$ allows to draw additional conclusions about $\boldsymbol{C_1}$. This is the case although the edge in the causal graph points from $\boldsymbol{C_1}$ to $\boldsymbol{X}$, i.e. $\boldsymbol{C_1}$ affects $\boldsymbol{X}$, not vice versa. For example, if $\boldsymbol{X}$ was simply the sum of $\boldsymbol{C_1}$ and the random term $\boldsymbol{U_X}$, a high value of $\boldsymbol{X}$ would probably imply a high value of $\boldsymbol{C_1}$. These conclusions about $\boldsymbol{C_1}$ cannot be drawn in the modified system, where the edge from $\boldsymbol{C_1}$ to $\boldsymbol{X}$ is removed. The knowledge about $\boldsymbol{C_1}$ allows to draw further conclusions about $\boldsymbol{Y}$ because $\boldsymbol{C_1}$ also affects $\boldsymbol{Y}$. Summarizing, due to the confounding influence of $\boldsymbol{C_1}$, knowing $\boldsymbol{X}$ reveals

more about $\boldsymbol{Y}$ in the original system than in the modified system, which is why the original expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}]$ and the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})]$ differ.

If we could observe the modified system, i.e. if we could experimentally set variable $\boldsymbol{X}$ to arbitrary values $\boldsymbol{x}$, we could approximate the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})]$ by training a suitable (see Sect. 2.2.1) statistical model on the observed tuples $(\boldsymbol{x},\boldsymbol{y})$ to predict $\boldsymbol{Y}$ given $\boldsymbol{X}$. However, in the cases considered in the proposed methodology, it is impossible or undesirable to experimentally set $\boldsymbol{X}$ to $\boldsymbol{x}$. Thus, we can only observe the original system and approximate the original expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}]$ by analogously training a statistical model on observed tuples $(\boldsymbol{x},\boldsymbol{y})$ of the original system. Consequently, we have to bridge the gap between the original expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}]$ and the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})]$.

2.1.2 Adjustment criteria

To bridge the gap between the original expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}]$ and the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})]$, we must take into account variables other than $\boldsymbol{X}$ and $\boldsymbol{Y}$. Indeed, in the example from Fig. 1, we showed that original and post-intervention expected values differ because, in the original system, knowing $\boldsymbol{X}$ allows inferences about $\boldsymbol{C_1}$ that are not possible in the modified system. However, if we actually knew $\boldsymbol{C_1}$, this would not be the case, thus, the original expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{C_1}=\boldsymbol{c_1}]$ and the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\boldsymbol{C_1}=\boldsymbol{c_1}]$ are identical. Analogously to $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}]$, the expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{C_1}=\boldsymbol{c_1}]$ can be approximated by observing the original system and training a statistical model on the observed tuples $(\boldsymbol{x},\boldsymbol{y},\boldsymbol{c_1})$ to predict $\boldsymbol{Y}$ given $\boldsymbol{X}$ and $\boldsymbol{C_1}$. Therefore, this equality allows to approximate the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\boldsymbol{C_1}=\boldsymbol{c_1}]$ by only observing the original system and *without* experimentally setting $\boldsymbol{X}$ to $\boldsymbol{x}$.

In the proposed methodology, we exploit the fact that the equality

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\{\boldsymbol{C_\ell}=\boldsymbol{c_\ell}\}_{\ell=1}^{k}]=\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C_\ell}=\boldsymbol{c_\ell}\}_{\ell=1}^{k}] \tag{4}$$

holds for any causal graph, thus allowing to determine the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C_\ell}=\boldsymbol{c_\ell}\}_{\ell=1}^{k}]$ from observations alone, if the additional variables $\boldsymbol{C_\ell}\in\mathbb{R}^{d_\ell},\ell=1,\ldots,k$, fulfil the following adjustment criteria (Shpitser et al., 2010):

1. The variables $\{\boldsymbol{C_\ell}\}_{\ell=1}^{k}$ block all non-causal paths from $\boldsymbol{X}$ to $\boldsymbol{Y}$ in the original causal graph.
2. No $\{\boldsymbol{C_\ell}\}_{\ell=1}^{k}$ lies on a causal path from $\boldsymbol{X}$ to $\boldsymbol{Y}$.

Here, a path is any consecutive sequence of edges. A path between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is causal from $\boldsymbol{X}$ to $\boldsymbol{Y}$ if all edges point towards $\boldsymbol{Y}$, and non-causal otherwise. A path is *blocked* by a set $S=\{\boldsymbol{C_\ell}\}_{\ell=1}^{k}$ of nodes if either (i) the path contains at least one edge-emitting node, i.e. a node with at least one adjacent edge pointing away from the node ($\ldots\leftrightarrow\boldsymbol{C}\rightarrow\ldots$), that is in S (e.g. the path $\boldsymbol{X}\leftarrow\boldsymbol{C_1}\rightarrow\boldsymbol{Y}$ in Fig. 1 is blocked by S if S contains $\boldsymbol{C_1}$); or (ii) the path contains at least one collision node, i.e. a node with both adjacent edges pointing towards the node ($\ldots\rightarrow\boldsymbol{C}\leftarrow\ldots$), which is outside S and has no descendants in S (e.g. the path $\boldsymbol{X}\rightarrow\boldsymbol{C}\leftarrow\boldsymbol{Y}$ is blocked if S does *not* contain $\boldsymbol{C}$).

The first adjustment criterion generalizes the example of $\boldsymbol{C_1}$ in Fig. 1, where adjusting for the *edge-emitting* node $\boldsymbol{C_1}$, i.e. considering $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{C_1}=\boldsymbol{c_1}]$ rather than $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}]$, blocks the non-causal path $\boldsymbol{X} \leftarrow \boldsymbol{C_1} \rightarrow \boldsymbol{Y}$ such that $\boldsymbol{X}$ is only used to draw conclusions about $\boldsymbol{Y}$ via the causal path $\boldsymbol{X} \rightarrow \boldsymbol{C_2} \rightarrow \boldsymbol{Y}$. In general, the criterion ensures that $\boldsymbol{X}$ is only used to draw conclusions about $\boldsymbol{Y}$ via causal paths from $\boldsymbol{X}$ to $\boldsymbol{Y}$ and not via any *non-causal* path between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

The second adjustment criterion ensures that no causal path from $\boldsymbol{X}$ to $\boldsymbol{Y}$ is blocked, such that the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C_\ell}=\boldsymbol{c_\ell}\}_{\ell=1}^k]$ actually reflects the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$. For example, considering the causal path $\boldsymbol{X} \rightarrow \boldsymbol{C_2} \rightarrow \boldsymbol{Y}$ in Fig. 1, $\boldsymbol{C_2}$ blocks the only causal path between $\boldsymbol{X}$ and $\boldsymbol{Y}$. Thus, $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\boldsymbol{C_2}=\boldsymbol{c_2}] = \mathbb{E}[\boldsymbol{Y}|\boldsymbol{C_2}=\boldsymbol{c_2}]$ would indicate that there is no causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$.

Summarizing this section, we can approximate the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C_\ell}=\boldsymbol{c_\ell}\}_{\ell=1}^k]$ from observations alone, if we can describe the considered system by a causal graph and find variables $\boldsymbol{C_\ell} \in \mathbb{R}^{d_\ell}, \ell = 1,\dots,k$ that fulfil the above adjustment criteria. Describing the system by a causal graph requires knowledge on which variables are relevant to the considered relation (represented by the nodes in the graph) and on the existence of causal dependencies between these variables (represented by the edges in the graph). Nevertheless, it does not require knowledge on the sign or strength of these dependencies, i.e. on the structural equations. Note that the parents of $\boldsymbol{X}$ in the causal graph always fulfil the adjustment criteria. In the proposed methodology, we exploit the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C_\ell}=\boldsymbol{c_\ell}\}_{\ell=1}^k]$ to determine the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ as detailed in Sect. 2.2.2.

2.2 Steps of the methodology

The proposed methodology is as follows: given a complex relation between two variables $\boldsymbol{X} \in \mathbb{R}^d$ and $\boldsymbol{Y} \in \mathbb{R}^n$, for example soil moisture-precipitation coupling, we train a *causal* deep learning (DL) model to predict $\boldsymbol{Y}$ given $\boldsymbol{X}$ and additional input variables $\boldsymbol{C_\ell} \in \mathbb{R}^{d_\ell}, \ell = 1,\dots,k$. In a second step, we perform a sensitivity analysis of the trained model to analyze how $\boldsymbol{Y}$ would change if we changed $\boldsymbol{X}$, i.e. to determine the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$.

2.2.1 Training a *causal* DL model

DL models (LeCun et al., 2015; Reichstein et al., 2019) learn statistical associations between their input and target variables. By training a causal DL model, we mean that we train a DL model that approximates for each input tuple $(\boldsymbol{x},\{\boldsymbol{c_\ell}\}_{\ell=1}^k)$ the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C_\ell}=\boldsymbol{c_\ell}\}_{\ell=1}^k]$, i.e. the model approximates the map

$$(\boldsymbol{x},\{\boldsymbol{c_\ell}\}_{\ell=1}^k) \rightarrow \mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C_\ell}=\boldsymbol{c_\ell}\}_{\ell=1}^k]. \tag{5}$$

To obtain a causal DL model, the loss function, model architecture and additional input variables $\{\boldsymbol{C_\ell}\}_{\ell=1}^k$ have to be chosen carefully. In particular, we choose a loss function that is minimized by the original expected value of $\boldsymbol{Y}$ given $\boldsymbol{X}$ and the other input variables, i.e. by the map

$$(\boldsymbol{x},\{\boldsymbol{c_\ell}\}_{\ell=1}^k) \rightarrow \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\{\boldsymbol{C_\ell}=\boldsymbol{c_\ell}\}_{\ell=1}^k]. \tag{6}$$

An example for such a loss function is the expected mean squared error,

$$(\boldsymbol{m} : (\boldsymbol{X}, \{\boldsymbol{C_\ell}\}_{\ell=1}^k) \to \mathbb{R}^n) \to \mathbb{E}[(\boldsymbol{Y} - \boldsymbol{m}(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k))^2], \tag{7}$$

which maps a function $\boldsymbol{m} : (\boldsymbol{X}, \{\boldsymbol{C_\ell}\}_{\ell=1}^k) \to \mathbb{R}^n$, representing the predictions of the DL model, to its expected mean squared error (Miller et al., 1993). Furthermore, in terms of model architecture, we choose a differentiable DL model (e.g. a neural network) that can represent the potentially complicated function from Eq. 6. Finally, we choose additional input variables $\{\boldsymbol{C_\ell}\}_{\ell=1}^k$ that fulfil the adjustment criteria from Sect. 2.1.2, such that the maps from Eq. 5 and Eq. 6 become identical. The choice of additional input variables requires prior knowledge on which variables are relevant for the considered relation, and on the existence of causal dependencies between these variables. However, it does not require prior knowledge on the strength, sign, or functional form of these dependencies (cf. Sect. 2.1.2), which can be obtained from the proposed methodology.

2.2.2 Sensitivity analysis of the trained model

To determine the causal effect of $\boldsymbol{X} \in \mathbb{R}^d$ on $\boldsymbol{Y} \in \mathbb{R}^n$, we consider partial derivatives of the map from Eq. 5, i.e.

$$s_{ij}(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k) = \frac{\partial \mathbb{E}[\boldsymbol{Y}_i | do(\boldsymbol{X} = \boldsymbol{x}), \{\boldsymbol{C_\ell} = \boldsymbol{c_\ell}\}_{\ell=1}^k]}{\partial \boldsymbol{X}_j}, \tag{8}$$

where $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, d\}$. These partial derivatives indicate how $\boldsymbol{Y}_i$ is expected to change if we experimentally varied the value of $\boldsymbol{X}_j$ by a small amount for given values $\boldsymbol{X} = \boldsymbol{x}, \{\boldsymbol{C_\ell} = \boldsymbol{c_\ell}\}_{\ell=1}^k$. We approximate these derivatives by the corresponding partial derivatives of the DL model, i.e. by the derivative of the predicted $\boldsymbol{Y}_i$ with respect to the input $\boldsymbol{X}_j$, denoted $q_{ij}(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k)$.

The target quantity in the proposed methodology is the expected value of $s_{ij}(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k)$ with respect to the probability distribution of $\boldsymbol{X}$ and $\{\boldsymbol{C_\ell} = \boldsymbol{c_\ell}\}_{\ell=1}^k$, i.e. $\overline{s_{ij}} = \mathbb{E}_{\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k}[s_{ij}(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k)]$. This quantity, which we refer to as the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$, indicates how $\boldsymbol{Y}_i$ is expected to change if we experimentally varied the value of $\boldsymbol{X}_j$ by a small amount. To approximate this quantity, we average the partial derivatives $q_{ij}(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k)$ of the DL model over a large number of observed tuples $(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k)$. For instance, when studying soil moisture-precipitation coupling, we average $q_{ij}(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k)$ over the T samples from the test set, i.e. we consider

$$\overline{q_{ij}} = \frac{1}{T} \sum_{(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k) \in \text{test set}} q_{ij}(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k). \tag{9}$$

Note that one might also combine partial derivatives for different tuples (i, j), for example to analyze the impact of a change in $\boldsymbol{X}_j$ on the sum $\sum_{i=1}^n \boldsymbol{Y}_i$. When studying soil moisture-precipitation coupling, we combine different partial derivatives to study the local and regional impact of soil moisture changes on precipitation (see Sect. 3.4).

In theory, the proposed methodology identifies the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ exactly. In practice, however, we make two important approximations. First, due to the complexity of the Earth system, the additional input variables $\{\boldsymbol{C_\ell}\}_{\ell=1}^k$ may not strictly fulfil the adjustment criteria from Sect. 2.1.2, such that the map from Eq. 6 is only approximately identical to the map from Eq. 5. Second, the DL model only approximates the map from Eq. 6. Thus, the partial derivatives $q_{ij}(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k)$ of

the DL model only approximate the partial derivatives $s_{ij}(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^{k})$ that we are interested in. We will come back to this in Sects. 3.3 and 4.

3 Application to soil moisture-precipitation coupling

As an illustrative example, we apply the proposed methodology to study soil moisture-precipitation coupling, i.e. the question how precipitation changes if soil moisture is changed. Although it is well-known that soil moisture affects precipitation (Seneviratne et al., 2010; Santanello et al., 2018), it remains unclear whether an increase in soil moisture results in an increase or decrease in precipitation. This is due to several concurring pathways of soil moisture-precipitation coupling (see Fig. 2). Improving our understanding of soil moisture-precipitation coupling is important to improve precipitation predictions with numerical models.

We apply the proposed methodology to study soil moisture-precipitation coupling across Europe at a short time scale of 3 to 4 hours. Namely, we train a causal DL model to predict precipitation $\boldsymbol{P}[t+4\text{ h}] \in \mathbb{R}^{80\times140}$ at 80×140 target pixels across Europe, given soil moisture $\boldsymbol{SM}[t] \in \mathbb{R}^{120\times180}$ and further input variables $\boldsymbol{C_\ell}[t] \in \mathbb{R}^{120\times180}$, e.g. antecedent precipitation, that approximately fulfil the adjustment criteria from Sect. 2.1.2, at 120×180 input pixels (see Fig. 3). In a second step, we perform a sensitivity analysis of the trained model to analyze how the precipitation predictions change if the soil moisture input variable is changed. Note, the input region is larger than the target region because $\boldsymbol{P}[t+4\text{ h}]$ depends on input variables in a surrounding region.

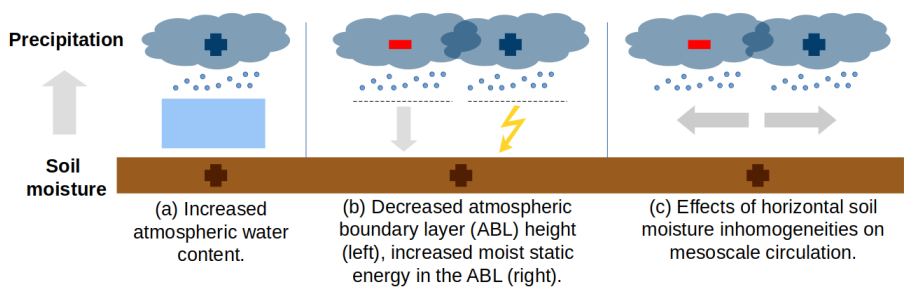

Figure 2. Concurring pathways of soil moisture-precipitation coupling. An increase in soil moisture can increase latent heat flux and decrease sensible heat flux at the land surface (Seneviratne et al., 2010). This can increase precipitation via an increase in atmospheric water content (a; Eltahir, 1998). At the same time, it can increase or decrease precipitation via boundary layer dynamics (b; Findell and Eltahir, 2003a, b; Gentine et al., 2013), or via effects of spatial heterogeneity in latent and sensible heat fluxes on mesoscale circulations (c; Eltahir, 1998; Adler et al., 2011; Taylor et al., 2011; Taylor, 2015).

3.1 Data

The data underlying our example are ERA5 hourly data (Hersbach et al., 2018) constituting an atmospheric reanalysis of the past decades (1950 to today) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Reanalysis

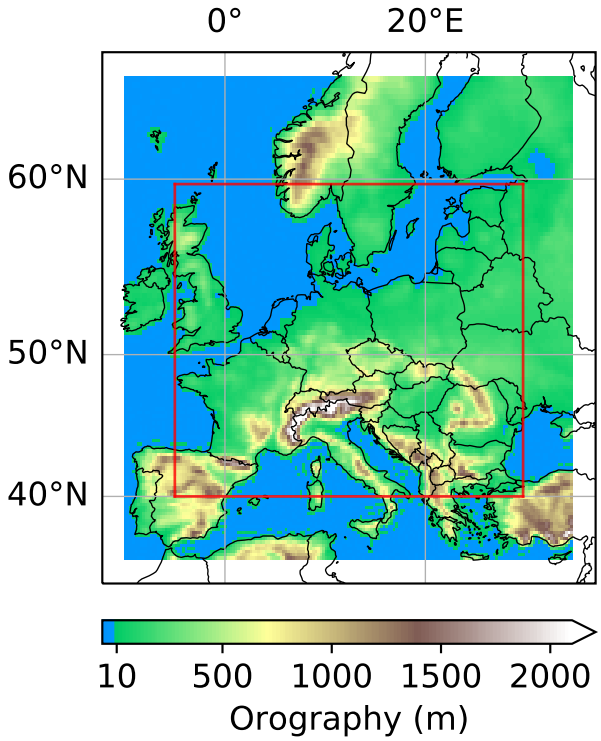

Figure 3. Input and target regions in the example of soil moisture-precipitation coupling. The colored region represents the 120×180 pixels input region, the red box the 80×140 pixels target region. Note that the offset between input and target region is 20 pixels on each side and distorted by the projection.

means simulation data and observations have been merged into a single description of the global climate and weather using data assimilation technologies. ERA5 data contain hourly estimates for a large number of atmospheric, ocean-wave and land-surface quantities on a regular latitude-longitude grid of 0.25 degrees (≈ 30 km). In this study, soil moisture refers to the ERA5 variable "volumetric soil water in the upper soil layer (0-7 cm)". The target variable, precipitation $\boldsymbol{P}[t+4\text{ h}]$, represents an accumulation of precipitation over the time interval $[t+3\text{ h}, t+4\text{ h}]$. In our analyses, we consider ERA5 data from 1979 to 2019 across Europe. Because soil moisture-precipitation coupling in Europe is strongest during the summer months, we only consider the months June, July and August. Further, we restrict our analyses to daytime processes considering precipitation predictions, $\boldsymbol{P}[t+4\text{ h}]$, for times $t+4$ h between noon and 11 pm UTC.

3.2 Loss function, model architecture and training

As described in Sect. 2.2.1, the loss function should be minimized by the expected value of precipitation $\boldsymbol{P}[t+4\text{ h}]$, given soil moisture $\boldsymbol{SM}[t]$ and the other input variables $\boldsymbol{C_\ell}[t]$, i.e. by the function (cf. Eq. 6)

$$(\boldsymbol{SM}[t], \{\boldsymbol{C_\ell}[t]\}_{\ell=1}^k) \rightarrow \mathbb{E}[\boldsymbol{P}[t+4\text{ h}]|\boldsymbol{SM}[t], \{\boldsymbol{C_\ell}[t]\}_{\ell=1}^k]. \tag{10}$$

This holds true for the expected mean squared error from Eq. 7. Given N training time steps t_i, associated values $(\boldsymbol{SM}[t_i], \{\boldsymbol{C_\ell}[t_i]\}_{\ell=1}^k, \boldsymbol{P}[t_i+4\text{ h}])_{i=1}^N$, and model predictions $\boldsymbol{m}(\boldsymbol{SM}[t_i], \{\boldsymbol{C_\ell}[t_i]\}_{\ell=1}^k)_{i=1}^N$, the expected mean squared error is approximated by the sum

$$\frac{1}{N}\sum_{i=1}^{N}\text{mean}((\boldsymbol{P}[t_i+4\text{ h}] - \boldsymbol{m}(\boldsymbol{SM}[t_i], \{\boldsymbol{C_\ell}[t_i]\}_{\ell=1}^k))^2). \tag{11}$$

Here, the mean operator denotes the mean over the 80×140 target pixels across Europe.

The employed DL model should be able to represent the presumably highly nonlinear function from Eq. 10. We choose a convolutional neural network (CNN; LeCun et al., 2015) whose architecture is inspired by the U-Net architecture (see Fig. 4; Ronneberger et al., 2015). Two concepts are important in applying CNNs in representing the function from Eq. 10. The first is the concept of receptive fields. Namely, the prediction of the model at some target location is fully determined by the input variables in a surrounding region, the so-called receptive field. In our case, the size of the receptive field is $\leq 52 \times 52$ pixels, i.e. the precipitation prediction at a target location is fully determined by the input variables in a $\leq 52 \times 52$ pixels surrounding region.

The second concept is that of translation invariance. Translation invariance means that the function $\hat{f}$, which maps the input variables in the receptive field to a prediction, is identical for all target locations. In our case, due to the arithmetic details of the considered model architecture (Dumoulin and Visin, 2016), the DL model is block translation invariant, i.e. the prediction at a target location (i, j) is not determined by a single function $\hat{f}$ for all target locations, but by one of 4×4 fixed functions $\hat{f}_{nk}, n, k = 1, \ldots, 4$ depending on the values $i \bmod 4$ and $j \bmod 4$.

Both concepts, receptive field and translation invariance, are important features of CNNs, because they counteract overfitting, i.e. making (nearly) perfect predictions on the training data but not generalizing to unseen data. However, both concepts constitute constraints that may prevent CNNs from representing the function from Eq. 10. Indeed, the translation invariance requires including additional input variables $\{C_\ell\}_{\ell=1}^k$ that lead to spatial variability in soil moisture-precipitation coupling. We will discuss this in Sect. 3.3. Note that we can mostly ignore the general constraint of receptive fields, because the lead time of the predictions is only 4 h and the receptive field is large enough to take into account all relations between soil moisture and precipitation at that time scale.

Before training the model, we split our data into training, validation and test sets. Due to potential correlations between subsequent time steps, an entirely random split would lead to high correlations between samples in training, validation and test sets. To achieve independence between samples belonging to different sets, we randomly choose all samples from the years 2010 and 2016 for validation, all samples from the years 2012 and 2018 for testing and all samples from the remaining 37 years for training. The test set is not used during the entire training and tuning process of the model.

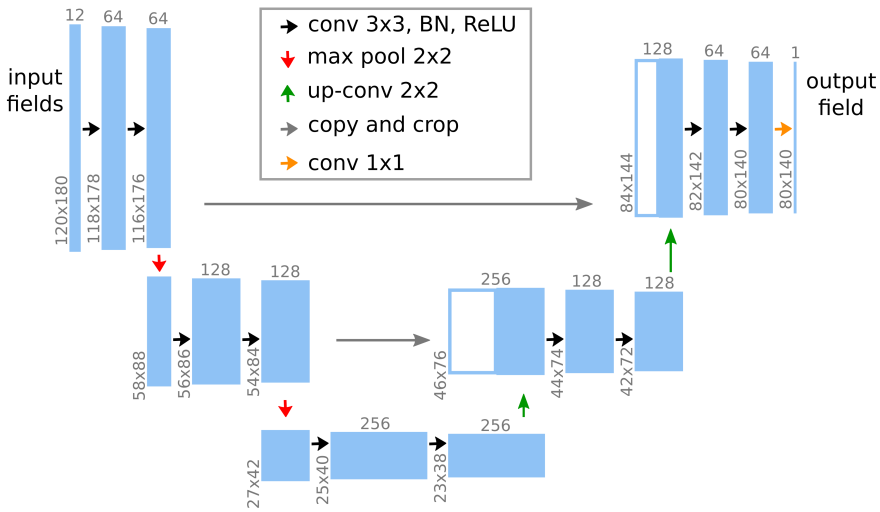

Figure 4. Model architecture in the example of soil moisture-precipitation coupling. The leftmost blue box represents the input to the model, which consists of 12 variables (including soil moisture) at the 120×180 input pixels (see Fig. 3). This input is passed through multiple sequential modules represented by the arrows. Each module performs simple mathematical operations on its respective inputs and produces an output that is fed to the next module. This output is represented by the next blue box and, in general, differs in shape from the input, as indicated by the grey upright and rotated numbers. For details on the mathematical operations we refer to (Ronneberger et al., 2015). The rightmost blue box represents the output of the model, which consists of the precipitation prediction at the 80×140 target pixels. The combination of multiple simple modules allows the model to represent complex functions.

During training, the Adam optimizer (Kingma and Ba, 2017) is used to adapt the approximately 2.3 million, randomly initialized weights of the model to minimize the mean squared error on the training set. In terms of implementation, we use the Pytorch (Paszke et al., 2019) wrapper skorch (Tietz et al., 2017) with default parameters for training the model, set the maximum number of epochs to 200; the learning rate in the Adam optimizer to $1e-3$; the batch size to 64; and patience for early stopping (i.e. the number of epochs after which training stops if the loss function evaluated on the validation set does not improve by some threshold) to 30 epochs. During training, we further use data augmentation. Namely, we randomly rotate by 180°(or not) and subsequently horizontally flip (or not) the considered region for each training sample and each training epoch independently. Similar to the translation invariance of the model, this requires including input variables which lead to spatial variability in soil moisture-precipitation coupling as discussed in the next section.

3.3 Choice of input variables

The choice of additional input variables $\{\boldsymbol{C_\ell}\}_{\ell=1}^{k}$ represents a crucial aspect of the proposed methodology for two reasons (cf. Sect. 2.2.2). First, we need the additional input variables to (approximately) fulfil the adjustment criteria from Sect. 2.1.2, such that the mapping of input variables $(\boldsymbol{SM}[t], \{\boldsymbol{C_\ell}[t]\}_{\ell=1}^{k})$ to $\mathbb{E}[\boldsymbol{P}[t+4\text{ h}]|\boldsymbol{SM}[t], \{\boldsymbol{C_\ell}[t]\}_{\ell=1}^{k}]$ (cf. Eq. 10) is a good

approximation of the map

$$(\boldsymbol{SM}[t], \{\boldsymbol{C_\ell}[t]\}_{\ell=1}^{k}) \rightarrow \mathbb{E}[\boldsymbol{P}[t+4\ \mathrm{h}]|do(\boldsymbol{SM}[t]), \{\boldsymbol{C_\ell}[t]\}_{\ell=1}^{k}]. \tag{12}$$

Second, the choice of additional input variables affects how accurately the CNN approximates the map from Eq. 10, and finally the partial derivatives of this map with respect to $\boldsymbol{SM}[t]$ that are computed in the sensitivity analysis (see Sect. 3.4).

Choosing additional input variables that fulfil the adjustment criteria is usually based on a causal graph of the considered system. However, a generally applicable causal graph of the Earth system does not exist. Thus, we make use of the fact that causal parents of $\boldsymbol{SM}[t]$ always fulfil the adjustment criteria, i.e. we look for a set of Earth system variables that is sufficient to determine $\boldsymbol{SM}[t]$ while not being affected by $\boldsymbol{SM}[t]$. Given the temporal resolution of the ERA5 data and the time scale of our analysis, a reasonable example is the set of variables in the second column in Fig. 5.

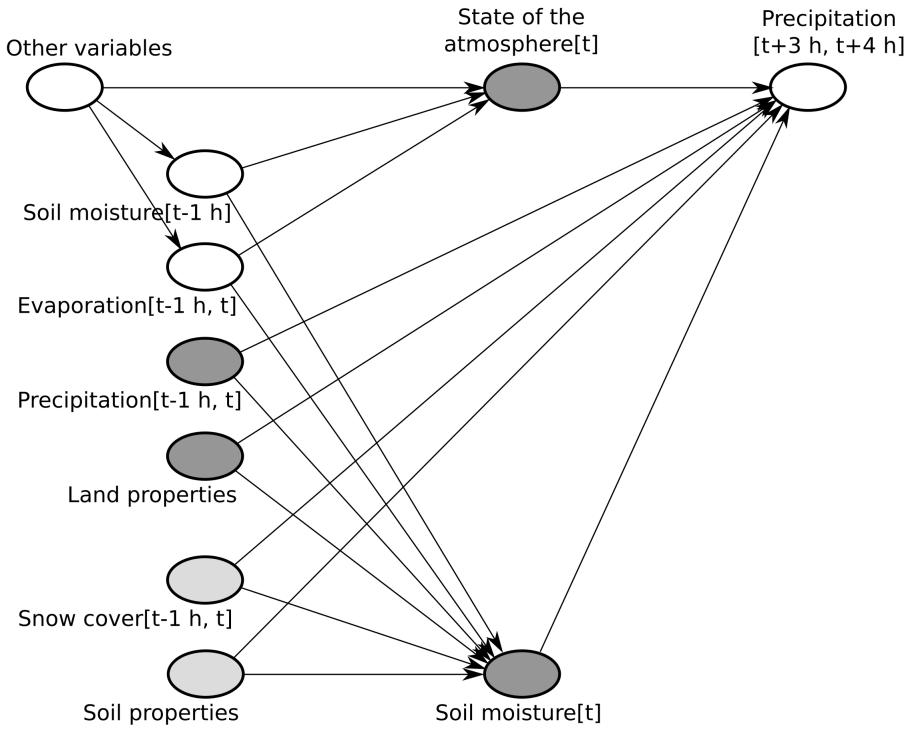

Figure 5. Causal graph in the example of soil moisture-precipitation coupling. The dark grey nodes represent the chosen input variables, while light grey nodes represent variables that are ignored in our analysis (see text). Land properties comprise the time-independent variables topography, land-sea mask, and fractions of high and low vegetation cover. The state of the atmosphere at time t is represented by temperature and dew point temperature at $2\ \mathrm{m}$ height at time t, as well as wind at $100\ \mathrm{m}$ height at time t. In addition to these variables, we included short- and long-wave radiation at the land surface at time t. Note that the depicted causal graph only includes nodes and edges that are relevant for the adjustment criteria from Sect. 2.1.2 (e.g. no edge from "other variables" to $\boldsymbol{P}[t-1\ \mathrm{h}, t]$, and no nodes on the causal path from $\boldsymbol{SM}[t]$ to $\boldsymbol{P}[t+3\ \mathrm{h}, t+4\ \mathrm{h}]$, such as evaporation$[t, t+3\ \mathrm{h}]$).

If we included all of these variables, the adjustment criteria would be met and the map from Eq. 10 would be identical to that from Eq. 12. Nevertheless, obtaining a good approximation of the map from Eq. 10 with our DL model would be difficult due to the strong correlation between $\boldsymbol{SM}[t-1\ \mathrm{h}]$ and $\boldsymbol{SM}[t]$. Furthermore, the strong correlation between evaporation$[t-1\ \mathrm{h}, t]$ and evaporation$[t, t+3\ \mathrm{h}]$ may prevent us from identifying any causal effect of $\boldsymbol{SM}[t]$ on $\boldsymbol{P}[t+4\ \mathrm{h}]$, because evaporation$[t, t+3\ \mathrm{h}]$ is a direct descendant of $\boldsymbol{SM}[t]$ on every causal path from $\boldsymbol{SM}[t]$ to $\boldsymbol{P}[t+4\ \mathrm{h}]$ (cf. motivation of the second adjustment criterion in Sect. 2.1.2). Therefore, we decided to exclude $\boldsymbol{SM}[t-1\ \mathrm{h}]$ and evaporation$[t-1\ \mathrm{h}, t]$. Nevertheless, this leads to unblocked non-causal paths between $\boldsymbol{SM}[t]$ and $\boldsymbol{P}[t+4\ \mathrm{h}]$ via these variables (e.g. $\boldsymbol{SM}[t] \leftarrow \boldsymbol{SM}[t-1\ \mathrm{h}] \rightarrow$ state of the atmosphere$[t] \rightarrow \boldsymbol{P}[t+4\ \mathrm{h}]$). To block these paths, we include additional input variables that represent the state of the atmosphere at time t.

Approximating the map from Eq. 10 and its partial derivatives with respect to $\boldsymbol{SM}[t]$ gets more difficult with increasing number of input variables. This is because additional input variables increase the complexity of this map, and the general risk of overfitting. Therefore, and because $\boldsymbol{SM}[t-1\ \mathrm{h}]$ and evaporation$[t, t-1\ \mathrm{h}]$ presumably affect the lower atmosphere more strongly than the higher atmosphere, we focus on variables representing the state of the lower atmosphere in this example.

The above considerations are valid for any model architecture and training procedure. In our example, we further must take into account the translation invariance of the considered DL model, and the rotation and flipping of the region used for data augmentation during the training procedure. Theoretically, in order to achieve invariance, the most accurate option is to include latitude-longitude information as additional input variables. However, if we did so, the DL model would have to learn a different mapping for each location (i, j), and data augmentation in form of flipping and rotation of the region would not be useful. Instead, we include short- and long-wave radiation at the land surface$[t]$. Thus, the above requirement is approximately fulfilled and the model does not have to learn a different mapping for each location, which presumably leads to it learning a better approximation of the map from Eq. 10.

The choice of input variables is where we insert prior knowledge in the proposed methodology (cf. Sect. 2.2.1). There is no unique choice of input variables, but several subjective decisions that have to be made. For example, above we could have started from a different set of causal parents, e.g. going not one but several hours into the past from time t, but at least theoretically that makes no difference (see Sect. 4). Starting from a set of causal parents and replacing variables strongly correlated with $\boldsymbol{X}$, as described above, seems to be a valid strategy for the choice of input variables, which is applicable to many relations in the Earth system besides soil moisture-precipitation coupling. It is in line with the fact that causal parents always fulfil the adjustment criteria, and with the general finding from causality research that input variables strongly correlated with $\boldsymbol{X}$ reduce the efficiency of statistical estimators of causal effects (Witte et al., 2020). The causal graph clearly conveys to other scientists the assumptions underlying a specific application of the proposed methodology.

3.4 Sensitivity analysis

Given our trained DL model, we consider different combinations of partial derivatives of the model to study the local and regional effects of soil moisture changes on precipitation (cf. Sect. 2.2.2). We define the causal effect of a soil moisture change at a pixel (i, j) on precipitation at the very same pixel as local effect or local soil moisture-precipitation coupling. Accordingly,

we consider for each pixel (i, j) in the target region the partial derivative

$$q_{ij}^{loc} = \frac{\partial \boldsymbol{p}_{ij}(\boldsymbol{SM}, \{\boldsymbol{C_\ell}\}_{\ell=1}^{k})}{\partial \boldsymbol{SM}_{ij}}, \tag{13}$$

where $\boldsymbol{p}_{ij}$ denotes the precipitation prediction of the DL model for pixel (i, j), and $\boldsymbol{SM}$ and $\{\boldsymbol{C_\ell}\}_{\ell=1}^{k}$ are the input variables to the model. We average these derivatives over all input samples $(\boldsymbol{SM}, \{\boldsymbol{C_\ell}\}_{\ell=1}^{k})$ from the test set denoted by $\overline{q^{loc}}_{ij}$.

Next to the local soil moisture-precipitation coupling, we define the regional effect or regional soil moisture-precipitation coupling as the causal effect of a soil moisture change at a pixel (i, j) on precipitation in the entire target region. Accordingly, we consider for each pixel (i, j) in the target region the sum of partial derivatives

$$q_{ij}^{reg} = \sum_{\hat{i}=1}^{80} \sum_{\hat{j}=1}^{140} \frac{\partial \boldsymbol{p}_{\hat{i}\hat{j}}(\boldsymbol{SM}, \{\boldsymbol{C_\ell}\}_{\ell=1}^{k})}{\partial \boldsymbol{SM}_{ij}}. \tag{14}$$

Note that most of the derivatives in the sum are zero, because e.g. a change in soil moisture in Great Britain at time t does not affect precipitation in Italy four hours later. Outside of a 52×52 pixels surrounding region, this is enforced by the architecture of the DL model (cf. Sect. 3.2), and inside of this region, it is learned during training of the model. As for local soil moisture-precipitation coupling, $\overline{q^{reg}}_{ij}$ denotes the average of q_{ij}^{reg} over all input samples from the test set.

To obtain robust results, we computed local and regional couplings for 10 instances of the DL model that were trained from different random weight initializations. Next, we averaged the obtained couplings ($\overline{q^{loc}}_{ij}$ and $\overline{q^{reg}}_{ij}$) over the 10 instances. The results are shown in Fig. 6. Notably, the difference in sign between positive local and negative regional impact demonstrates the importance of taking into account non-local effects of soil moisture-precipitation coupling, which are neglected by many other approaches. Moreover, Fig. 6 indicates particularly strong local and regional couplings in mountainous regions and ridges. We will further discuss the correctness of these results in Sect. 4.

3.5 Comparison to other approaches

A common approach for studying relations in the Earth system is to consider the linear correlation between variables (Froidevaux et al., 2014; Welty and Zeng, 2018; Holgate et al., 2019). Here, we compare our results on regional soil moisture-precipitation coupling to results obtained from a linear correlation analysis. For each location in the considered target region, Fig. 7 shows the linear correlation coefficient of soil moisture $\boldsymbol{SM}[t]$ at that location and subsequent precipitation $\boldsymbol{P}[t + 4\ \mathrm{h}]$ summed over the 15×15 pixels surrounding region. In contrast to our analysis of regional soil moisture-precipitation coupling, the linear correlation analysis assumes linearity of relations between local soil moisture and regional precipitation, and neglects the difference between causality and correlation. The obtained regional soil moisture-precipitation "coupling" in Fig. 7 then also differs in sign and spatial pattern from the coupling in the right panel of Fig. 6, stressing the importance of accounting for nonlinear effects and for the difference between causality and correlation in the proposed methodology.

Another approach for studying soil moisture-precipitation coupling is to perform multiple numerical simulations that differ only in initial soil moisture and to analyze the differences in precipitation between these simulations (Imamovic et al., 2017; Baur et al., 2018; Leutwyler et al., 2021). This approach allows to evaluate the effects of soil moisture changes on precipitation

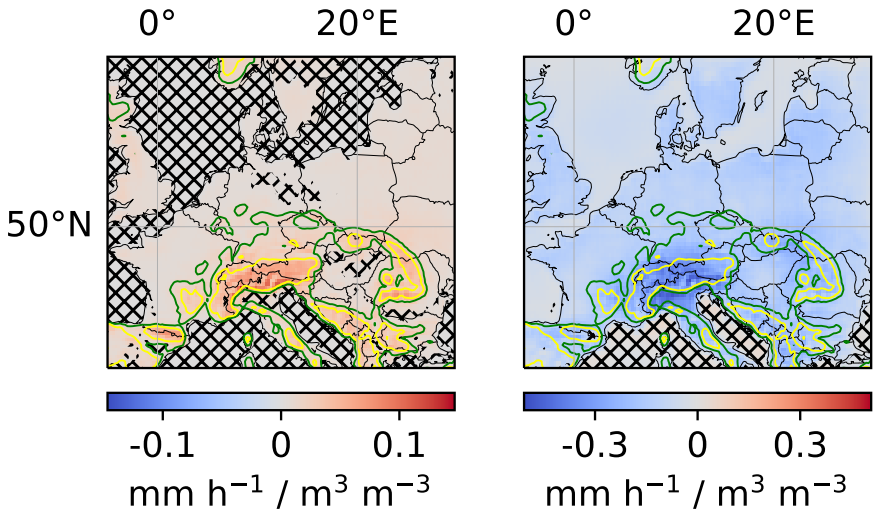

Figure 6. Local and regional soil moisture-precipitation couplings. Left: Impact of local soil moisture changes (m^3 water $\cdot\, \mathrm{m}^{-3}$ soil) on local precipitation ($\mathrm{mm\,h}^{-1}$) for each pixel in the target region (in the text denoted by $\overline{q^{loc}}_{ij}$). Right: Impact of local soil moisture changes on regional precipitation for each pixel in the target region (in the text denoted by $\overline{q^{reg}}_{ij}$). For better comparability of local and regional values, the unit $\mathrm{mm\,h}^{-1}$ for precipitation refers to a single pixel in both panels. Missing hatching indicates that the coupling reflects more than random correlations between soil moisture and precipitation in the training data, artifacts of the DL training procedure, seasonality, and the correlation between soil moisture and topography (see Sect. 4.2). The green and yellow elevation contour lines indicate 370 m and 750 m, respectively.

within the employed numerical model precisely. However, for some questions, it is computationally infeasible. For instance, in this work, we used ERA5 data to analyze the effects of soil moisture changes at each of 120×80 target pixels on subsequent precipitation in the target region. We averaged these effects over all time steps in two test years, constituting $2\,208$ time steps. Performing an analogous study based on numerical simulations would require at least $120 \cdot 80 \cdot 2\,208 = 21\,196\,800$ 4-hourly simulations with the ECMWF Earth system model used to produce the considered ERA5 data. Each simulation would be initialized with the state of the reference simulation at one of the $2\,208$ considered time steps, the only difference being that soil moisture would be slightly increased or decreased at one of the 120×80 target pixels. This corresponds to simulating approximately $10\,000$ years with the ECMWF Earth system model and is computationally infeasible. Furthermore, an advantage of the proposed methodology over approaches based on numerical simulations is that it can directly be applied to observational data, if suitable observational data are available. In this case, the proposed methodology does not rely on any assumptions incorporated into numerical models.

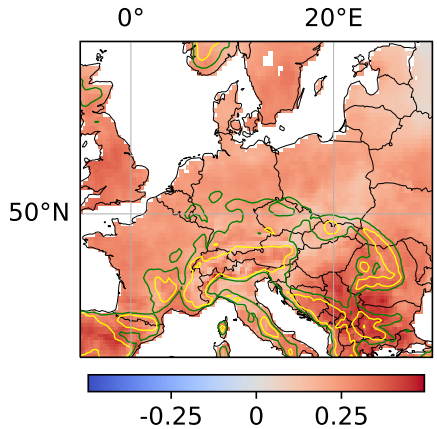

Figure 7. Linear correlation coefficient of local soil moisture and regional precipitation. For each location, the linear correlation coefficient of soil moisture $\boldsymbol{SM}[t]$ at the location and subsequent precipitation $\boldsymbol{P}[t+4\text{ h}]$ summed over the 15×15 pixels surrounding region of the location is shown.

4 Additional analyses to verify the results

To ensure that the proposed methodology provides reliable results, this section presents some additional analyses. Theoretically, the proposed methodology determines the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ exactly. However, in practice, we make two important approximations (cf. Sect. 2.2.2). First, the additional input variables $\{\boldsymbol{C_\ell}\}_{\ell=1}^{k}$ may not strictly fulfil the adjustment criteria from Sect. 2.1.2, such that the mapping of input variables to the original expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\{\boldsymbol{C_\ell}=\boldsymbol{c_\ell}\}_{\ell=1}^{k}]$ in Eq. 6 is only approximately identical to the mapping to the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C_\ell}=\boldsymbol{c_\ell}\}_{\ell=1}^{k}]$ in Eq. 5. Second, the DL model represents only an approximation of the map from Eq. 6. Both errors are difficult to quantify, because both maps are unknown.

For example, the performance of the DL model on the test set cannot indicate how well the DL model approximates the map from Eq. 6, because the loss value for this map is not known. For instance, for a system described by the causal graph $\boldsymbol{X} \rightarrow \boldsymbol{Y} \leftarrow \boldsymbol{C}$ and the structural equation $\boldsymbol{Y} = \boldsymbol{X} + 1000 \cdot \boldsymbol{C}$ (where $\boldsymbol{X}$ and $\boldsymbol{C}$ vary in similar ranges), the adjustment criteria from Sect. 2.1.2 imply that it suffices to consider $\boldsymbol{X}$ as input variable in the proposed methodology. Nevertheless, even if the trained DL model perfectly represented the map $\boldsymbol{x} \rightarrow \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}]$, the associated loss value would be high as knowing $\boldsymbol{X}$ does not reveal much about $\boldsymbol{Y}$, which is mainly determined by $\boldsymbol{C}$.

The results of the proposed methodology are the partial derivatives $\overline{q_{ij}}$ of the DL model computed in the sensitivity analysis. Due to the above approximations, these derivatives are only approximations of the partial derivatives $\overline{s_{ij}}$ of the map from Eq. 5, which represent the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ (cf. Sect. 2.2.2). However, even quantifying the two approximation errors mentioned above would not give us a good estimate of the errors in these results. In this section, we propose several additional analyses to build confidence in results obtained with the proposed methodology. Particularly, the proposed analyses show if

results are statistically significant, i.e. reflect more than random correlations or artifacts of the DL training procedure (Sect. 4.1), and if they reflect more than specific (known) correlations (Sect. 4.2). Moreover, the analyses proposed in Sect. 4.3 allow to identify (potentially unknown) spurious correlations in the results. Finally, we propose some further sanity checks in Sect. 4.4. We illustrate the analyses with our results on soil moisture-precipitation coupling from Sect. 3.

For reference only, we provide here the normalized mean squared error on the test set (target variable normalized to mean of 0 and standard deviation of 1 on the training set) for our application to soil moisture-precipitation coupling: it is 0.60 for the DL model. For a persistence prediction, i.e. when predicting the input $\boldsymbol{P}[t]$ as target $\boldsymbol{P}[t+4\text{ h}]$, which is a simple baseline prediction, it is 1.54.

4.1 Statistical significance

To test whether results obtained with the proposed methodology are statistically significant, i.e. represent more than random correlations between $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the training data and random artifacts of the procedure for training the DL model, we propose the following procedure. First, randomly permute $\boldsymbol{X}$ in the training data, thereby breaking all non-random correlations between $\boldsymbol{X}$ and $\boldsymbol{Y}$. For example, in the application to soil moisture-precipitation coupling, permute soil moisture temporally and spatially. Next, train a separate instance of the original DL model with a random initialization of model weights on the modified training data. Repeat this procedure several times. If the original results deviate significantly from the results obtained from this procedure, they are statistically significant.

Formally, we propose to interpret a result $r \in \mathbb{R}$ of the proposed methodology, e.g. local or regional soil moisture-precipitation coupling at some pixel (i, j) (cf. Sect. 3.4), as a sample of a random variable $\hat{r} : \Omega \to \mathbb{R}$, where Ω is the probability space

$$\Omega = \{\text{Training data}\} \times \{\text{Weight initialization of the DL model}\}. \tag{15}$$

Thus, $\hat{r}$ computes the considered result, e.g. local or regional soil moisture-precipitation coupling at pixel (i, j) according to the proposed methodology, for any given sample $\omega \in \Omega$. We define the null hypothesis that r represents random correlations between $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the training data, or random artifacts of the procedure for training the DL model. To test this hypothesis, we create m samples $\omega_0^1, \ldots, \omega_0^m$ of Ω by the above described procedure of permuting $\boldsymbol{X}$ and randomly initializing the weights of separate instances of the considered DL model. Moreover, we compute the associated values $r_0^i = \hat{r}(\omega_0^i), i = 1, \ldots, m$, representing samples of $\hat{r}$ under the null hypothesis.

If the original value r differs from these samples, we can reject the null hypothesis and conclude that r is statistically significant. In particular, if m is large enough, we can reject the null hypothesis at some significance level α (e.g. $\alpha = 5\ \%$), if the original value r lies outside the middle $100\ \% - \alpha$ of the values $r_0^1, \ldots, r_0^m$, i.e. if

$$r \notin [\text{percentile}(\{r_0^1, \ldots, r_0^m\}, \alpha/2), \text{percentile}(\{r_0^1, \ldots, r_0^m\}, 100\ \% - \alpha/2)]. \tag{16}$$

However, because we have to train m DL models for this analysis, it may not be feasible to choose m large enough to get reasonable approximations of these percentiles. In this case, we propose computing the mean μ and standard deviation σ of the values $r_0^1, \ldots, r_0^m$, *assuming* a normal distribution of $\hat{r}$ under the null hypothesis, and rejecting the null hypothesis at

significance level α if r is not in the middle $100\,\% - \alpha$ of the distribution $N(\mu, \sigma)$, i.e. if

$$r \notin [\text{percentile}(N(\mu,\sigma), \alpha/2), \text{percentile}(N(\mu,\sigma), 100\,\% - \alpha/2)]. \tag{17}$$

4.2 Known spurious correlations

As mentioned above, the proposed methodology identifies the exact causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ in theory, but not necessarily in practice, where results might reflect spurious correlations. In this section, we propose two analyses to test whether results obtained with the proposed methodology represent more than spurious correlations. The analyses apply whenever the spurious correlations are known, and $\boldsymbol{X}$ can be permuted such that the considered correlations are preserved while other correlations between $\boldsymbol{X}$ and $\boldsymbol{Y}$ break. For example, there exists a spurious correlation between $\boldsymbol{SM}[t]$ and $\boldsymbol{P}[t + 4\text{ h}]$ via topography, because topography affects both $\boldsymbol{SM}[t]$ and $\boldsymbol{P}[t + 4\text{ h}]$ ($\boldsymbol{SM}[t] \leftarrow$ topography $\rightarrow \boldsymbol{P}[t + 4\text{ h}]$, cf. Sect. 2.1.1). Further, there might exist a spurious correlation between $\boldsymbol{SM}[t]$ and $\boldsymbol{P}[t + 4\text{ h}]$ via seasonality, e.g. if both soil moisture and precipitation were generally lower in August than in June. Both correlations are preserved if we permute soil moisture year-wise as illustrated in Fig. 8. All other cases of spurious correlations are discussed in the next section, in particular unknown spurious correlations.

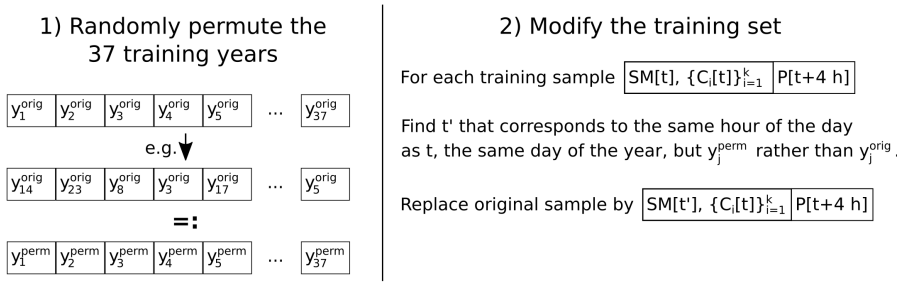

Figure 8. Modification of the training data for the year-wise permutation of $\boldsymbol{SM}[t]$. The modification of the test data works analogously.

The first proposed analysis is identical to the analysis described in Sect. 4.1 except that $\boldsymbol{X}$ in the training data is not permuted randomly, but such that the considered spurious correlations are preserved. If the original results deviate significantly from the results obtained in this analysis, they are statistically significant and do not only represent the considered spurious correlations. In our example of soil moisture-precipitation coupling, we permuted $\boldsymbol{SM}[t]$ year-wise as illustrated in Fig. 8 and trained $m = 10$ separate instances of the DL model. The analysis indicates that our results on soil moisture-precipitation coupling are statistically significant and represent more than correlations between soil moisture and topography or seasonality (missing hatching in Fig. 6). Intriguingly, the regional coupling is statistically significant (albeit weak) at most ocean locations, although one would not expect the DL model to learn a systematic effect of soil moisture variations on precipitation at these locations, since soil moisture does not vary at these locations. Indeed, we set soil moisture to 1 m^3 water per m^3 at all ocean locations for all time steps, while it is smaller than 0.75 at all non-ocean locations. We assume that the statistical significance of the regional coupling at ocean locations is an artifact of the DL model architecture, which favours generalization between locations, ocean and non-ocean.

The second proposed analysis evaluates whether the original DL model learned useful information in terms of predictive performance on the relation between $\boldsymbol{X}$ and $\boldsymbol{Y}$, apart from the considered spurious correlations. In the analysis, we train m separate instances of the original DL model on the original training data. The m instances differ in the random initialization of model weights (cf. Sect. 3.4). For each model instance, we compute the value of the loss function on the test set, obtaining m values $l_1, \ldots, l_m \in \mathbb{R}$. Next, for each model instance separately, we randomly permute $\boldsymbol{X}$ in the test data such that the considered spurious correlations are preserved, and compute the value of the loss function on the modified test set, obtaining m values $l_1^{\text{perm}}, \ldots, l_m^{\text{perm}} \in \mathbb{R}$. Finally, we use a permutation test (Hesterberg, 2014) to test if the expected value of the loss function is smaller on the original test set than on the modified test set. If this is the case, the DL models learned something useful in terms of predictive performance on the relation between $\boldsymbol{X}$ and $\boldsymbol{Y}$, apart from the considered spurious correlations. In our example of soil moisture-precipitation coupling, we trained $m = 10$ separate instances of the DL model. We considered the year-wise permutation of soil moisture in the test data as described above. In this case, the analysis indicates at a confidence level of 99 % that the model learned useful information in terms of predictive performance on soil moisture-precipitation coupling, apart from the correlations between soil moisture and topography or seasonality. However, for the validity of this analysis, it may be limiting that there are only two test years in this example and thus only one possible permutation of years apart from the original one. Therefore, we repeated the analysis and permuted soil moisture in the test data completely randomly in time. While this does not preserve correlations between soil moisture and seasonality, it still preserves the correlation between soil moisture and topography. Furthermore, it ensures the validity of the analysis as there are a lot of possible permutations. In this case, the analysis indicates at a confidence level of 99 % that the model learned useful information in terms of predictive performance on soil moisture-precipitation coupling, apart from the correlation between soil moisture and topography. Note that even if the first analysis indicates that some result reflects more than the considered correlations, it cannot exclude that the results are partly affected by the considered spurious correlations. Analogously, if the second analysis indicates that the DL model learned useful information in terms of predictive performance on the relation between $\boldsymbol{X}$ and $\boldsymbol{Y}$, apart from the considered spurious correlations, it cannot exclude that the predictions are partly affected by the considered spurious correlations.

4.3 Further spurious correlations

In the previous section, we analysed specific spurious correlations, i.e. spurious correlations that were known, and for that $\boldsymbol{X}$ could be permuted such that the spurious correlations are preserved, while other correlations between $\boldsymbol{X}$ and $\boldsymbol{Y}$ break. As an additional analysis to identify any spurious correlations reflected in obtained results, we propose a variant approach. The concept of the approach is related to the ideas in (Tesch et al., 2021) and (Peters et al., 2016). It consists of training separate instances of the original DL model (referred to as variant models) on modified prediction tasks (referred to as variant tasks) for which it is assumed that causal relations between input and target variables either remain stable or vary in specific ways. Subsequently, the results obtained from original and variant models are compared and it is evaluated whether they reflect the assumed stability or specific variations, respectively, of causal relations. If not, the original model or one of the variant models (or all models) learned spurious correlations.

For example, we may assume that the general (causal) mechanisms of soil moisture-precipitation coupling do not vary in time or space. Then, if the couplings in Fig. 6 reflect the causal effect of soil moisture on precipitation, we should obtain the same couplings from separate instances of the DL model that are trained only on

- data from the first or second half of the training years,
- data from June, July or August, or
- the left or right half of the considered region.

On the other hand, if Fig. 6 reflected spurious correlations *and* these spurious correlations differed for the different subsets of training data listed above, we should obtain different couplings from the different model instances.

Appendix Figs. A1 to A3 show the local and regional couplings obtained from the different model instances trained on the listed training subsets. As expected for the case that all instances learned the causal effect of soil moisture on precipitation, all couplings are very similar to the ones shown in Fig. 6. Note however that this does not guarantee that they show causal relations.

4.4 Task-specific sanity checks

To further assess the correctness and increase trust in results obtained from the proposed methodology, we propose to perform further, task-specific sanity checks. For instance, in our example of soil moisture-precipitation coupling, precipitation $\boldsymbol{P}$ can be partitioned into convective precipitation $\boldsymbol{P_{con}}$ (occurring at spatial scales smaller than the spatial resolution of the numerical model) and large-scale precipitation $\boldsymbol{P_{ls}}$ (occurring at larger spatial scales), such that $\boldsymbol{P} = \boldsymbol{P_{con}} + \boldsymbol{P_{ls}}$. Accordingly, soil moisture-precipitation coupling, $\boldsymbol{SM}$-$\boldsymbol{P}$ coupling, can be decomposed into the sum of $\boldsymbol{SM}$-$\boldsymbol{P_{con}}$ coupling and $\boldsymbol{SM}$-$\boldsymbol{P_{ls}}$ coupling. As a sanity check for the results in Fig. 6, we applied the proposed methodology to obtain $\boldsymbol{SM}$-$\boldsymbol{P_{con}}$ coupling and $\boldsymbol{SM}$-$\boldsymbol{P_{ls}}$ coupling by replacing $\boldsymbol{P}$ by $\boldsymbol{P_{con}}$ and $\boldsymbol{P_{ls}}$, respectively, and compared the sum of the obtained couplings with Fig. 6. Appendix Fig. A5 shows the sum of local and regional $\boldsymbol{SM}$-$\boldsymbol{P_{con}}$ and $\boldsymbol{SM}$-$\boldsymbol{P_{ls}}$ couplings, which are indeed very similar to the couplings shown in Fig. 6.

Further, $\boldsymbol{SM}$-$\boldsymbol{P}$ coupling can approximately be factorized into instantaneous (local) soil moisture-evaporation ($\boldsymbol{SM}$-$\boldsymbol{E}$) coupling times evaporation-precipitation ($\boldsymbol{E}$-$\boldsymbol{P}$) coupling. As another sanity check for the results in Fig. 6, we applied the proposed methodology to obtain $\boldsymbol{SM}$-$\boldsymbol{E}$ coupling and $\boldsymbol{E}$-$\boldsymbol{P}$ coupling by once replacing the target variable $\boldsymbol{P}$ by $\boldsymbol{E}$ and the other time replacing the input variable $\boldsymbol{SM}$ by $\boldsymbol{E}$. Appendix Fig. A7 shows the product of local $\boldsymbol{SM}$-$\boldsymbol{E}$ and local and regional $\boldsymbol{E}$-$\boldsymbol{P}$ couplings. The obtained couplings are very similar to the couplings shown in Fig. 6, despite being slightly weaker in general and far weaker in the high Alps.

4.5 Control experiment

As a simple control experiment for the proposed methodology and analyses, we replaced the target variable $\boldsymbol{P}[t+4\ \mathrm{h}]$ by random noise. As expected from the missing correlations between $\boldsymbol{SM}[t]$ and random noise, the methodology identified no statistically significant (cf. Sect. 4.1) causal effect of soil moisture on the target variable in this case.

Defining a more complex control experiment confirming the results obtained in the application to soil moisture-precipitation coupling is not possible. This is because the maps in Eq. 6 and Eq. 5, and thus the errors in their approximations, would differ if, for example, we replaced $\boldsymbol{SM}[t]$ by a variable $\boldsymbol{X}$ that is highly correlated with $\boldsymbol{P}[t+4\ \mathrm{h}]$ but does not causally affect $\boldsymbol{P}[t+4\ \mathrm{h}]$. However, we believe that the analyses proposed in this section build high confidence in the methodology and the results.

5 Conclusions

In this study, we proposed a novel methodology for studying complex, e.g. nonlinear and non-local, relations in the Earth system. The methodology is based on the recent idea of training and analyzing a DL model to gain new scientific insights into the relations between input and target variables. It extends this idea by combining it with concepts from causality research. A crucial aspect in the proposed methodology is the choice of additional input variables for the DL model. This choice requires prior knowledge on which variables are relevant to the considered relation, and on the existence of dependencies between these variables. However, it does not require prior knowledge on the strength or sign of these dependencies, which can be obtained from the proposed methodology. When the required prior knowledge does not exist, methods from causal discovery (Guo et al., 2021) might be used to identify a causal graph anyway, such that the proposed methodology might still be applicable.

In addition to the methodology, we presented analyses to assess whether results obtained with the proposed methodology are statistically significant, i.e. reflect more than random correlations or artifacts of the DL training procedure, whether they reflect more than specific (known) correlations, and whether they actually reflect causal rather than (potentially unknown) spurious correlations. Finally, we proposed sanity checks for the obtained results. While the analyses cannot guarantee the correctness of obtained results, we believe that the proposed analyses provide a solid indication of the correctness of obtained results. Taking into account the difference between causality and correlation, and overcoming common assumptions on linearity and locality in statistical approaches, as well as high computational costs and assumptions of numerical approaches, we believe that the proposed methodology may yield new scientific insights into various complex mechanisms in the Earth system.

As an illustrating example, we applied the methodology and the proposed analyses to study soil moisture-precipitation coupling in ERA5 climate reanalysis data across Europe. Our main findings are the difference in sign between positive local and negative regional impact and particularly strong local and regional couplings in mountainous regions and ridges. While we believe that these findings may contribute to a better understanding of soil moisture-precipitation coupling, in this article, we focused on demonstrating the methodology. An extension and discussion of our results on soil moisture-precipitation coupling in terms of physical processes are subject of a future study.

Code and data availability. The ERA5 climate reanalysis data (Hersbach et al., 2018) underlying this study are publicly available. Code to reproduce the study can be found here: https://doi.org/10.5281/zenodo.6385040.

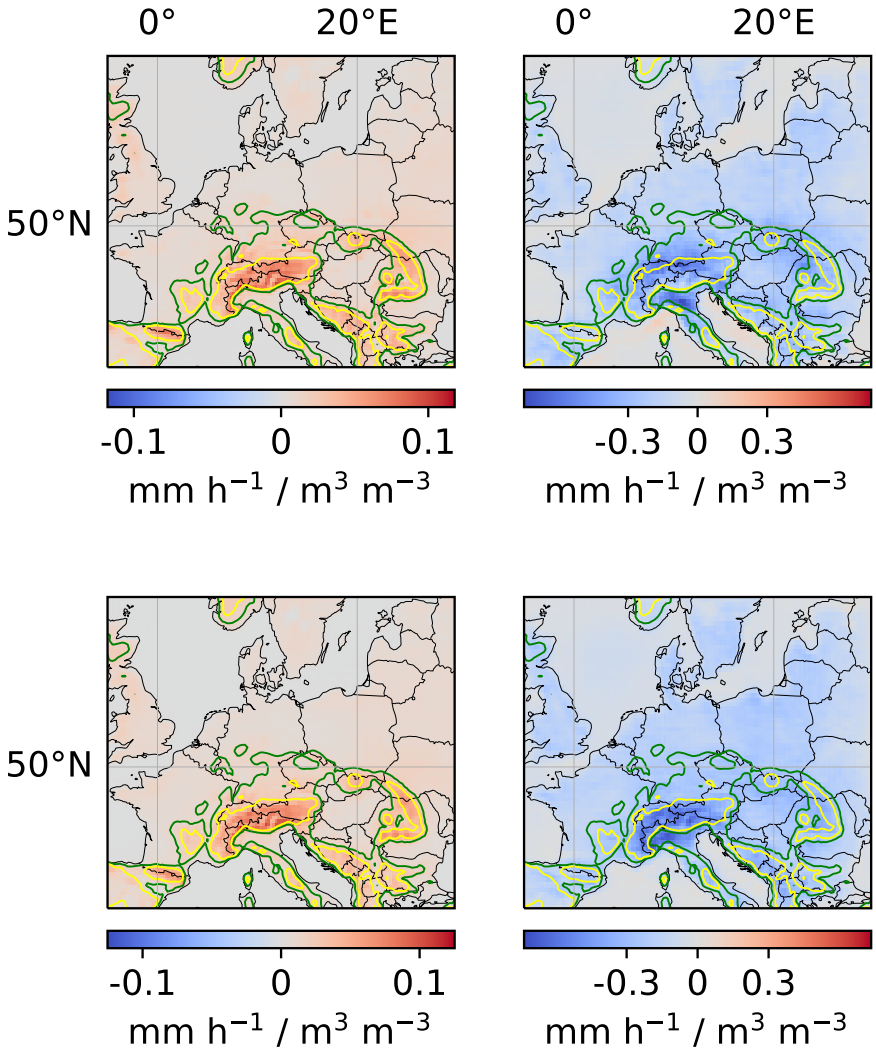

Figure A1. Local and regional soil moisture-precipitation couplings for models trained on the first and second half of the training years, respectively. Left column: local couplings. Right column: regional couplings. Upper row: model trained on the first half of all training years (1979-1997). Bottom row: model trained on the second half of all training years (1998-2019).

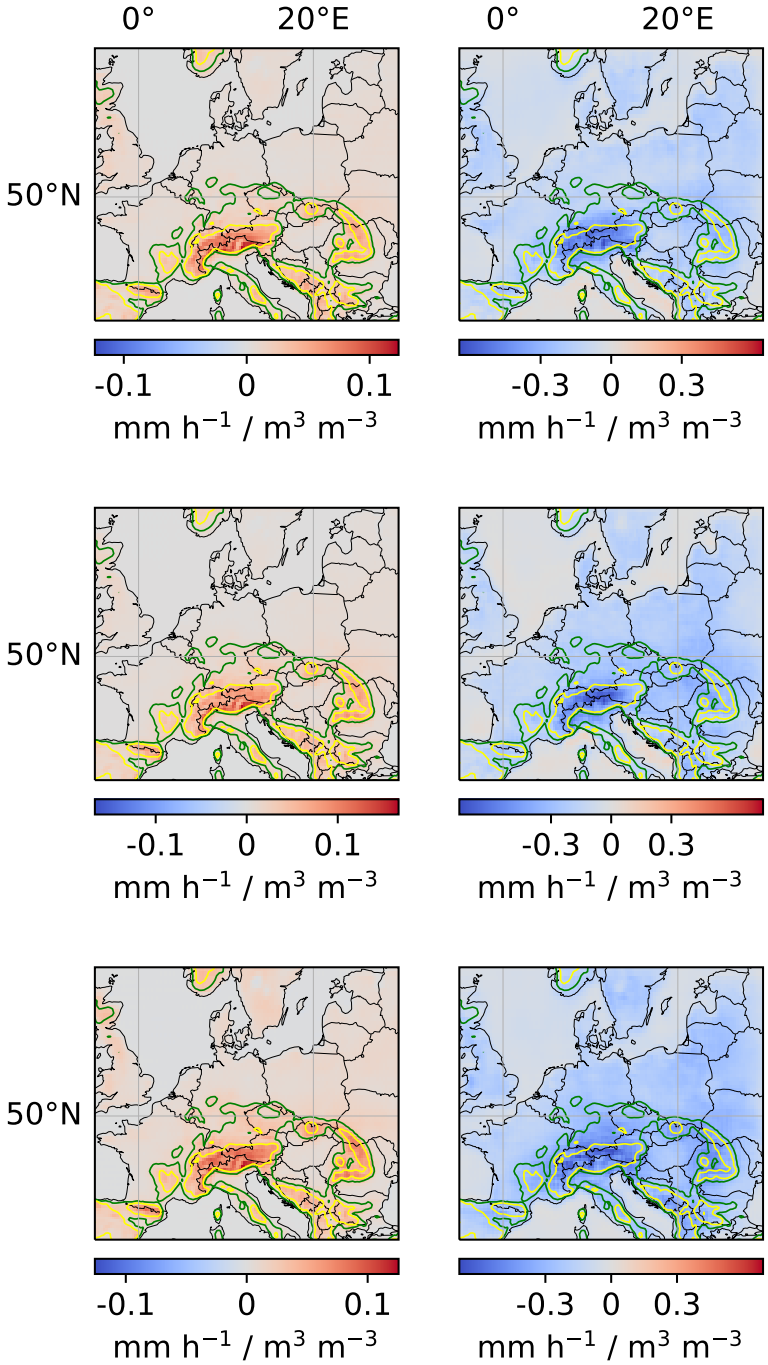

Figure A2. Local and regional soil moisture-precipitation couplings for models trained only on data from June, July and August, respectively. Left column: local couplings. Right column: regional couplings. Upper row: model trained on data from June. Centre row: model trained on data from July. Bottom row: model trained on data from August.

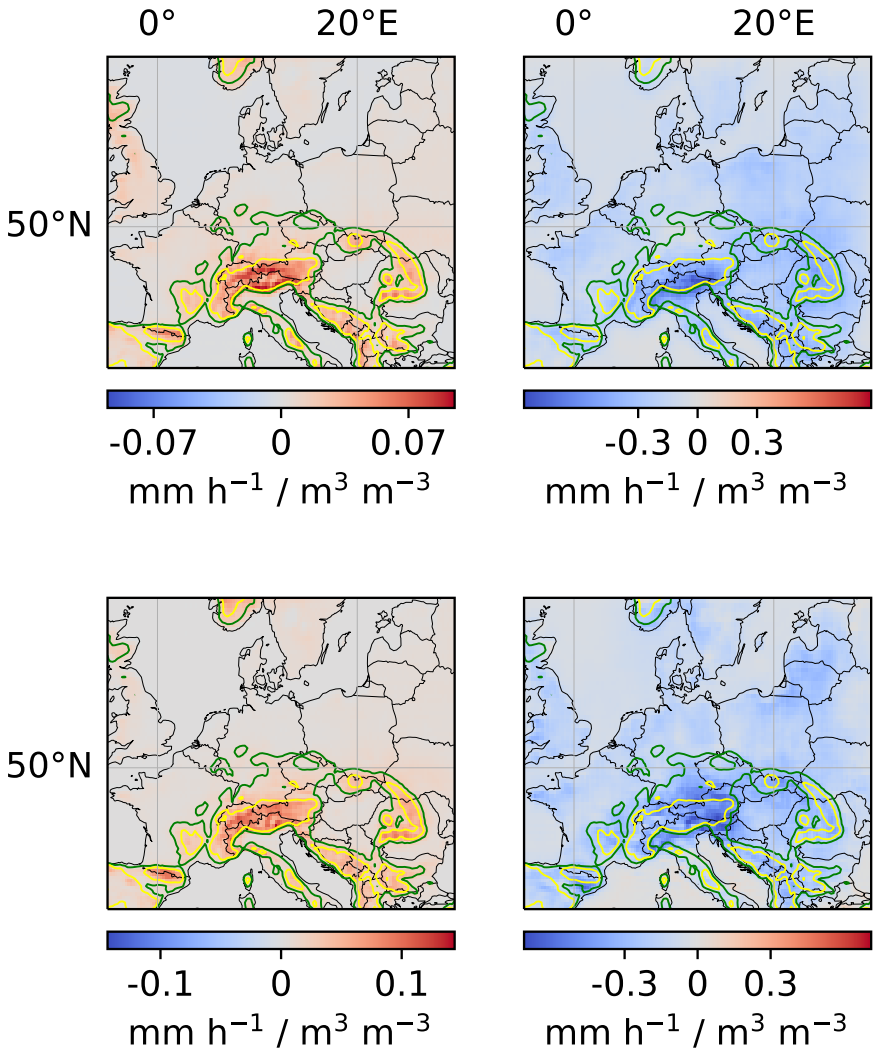

Figure A3. Local and regional soil moisture-precipitation couplings for models trained on the left and right half of the considered region, respectively. Left column: local couplings. Right column: regional couplings. Upper row: model trained on the left half of the considered region. Bottom row: model trained on the right half of the considered region (see Appendix Fig. A4). Note that the models were trained only on the left and right half, respectively, but the model architecture allows to compute local and regional couplings for the entire region.

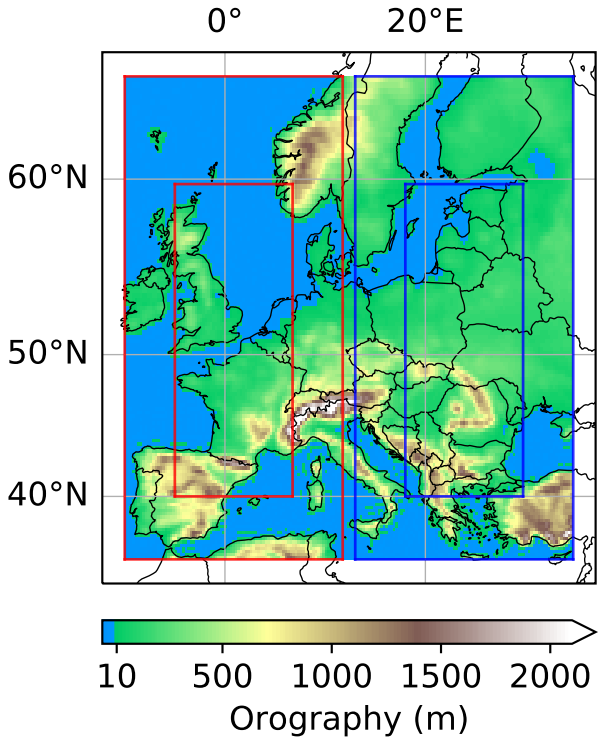

Figure A4. Location variant tasks. The input region was divided in a left and a right input region with corresponding target regions (indicated by the red and blue boxes).

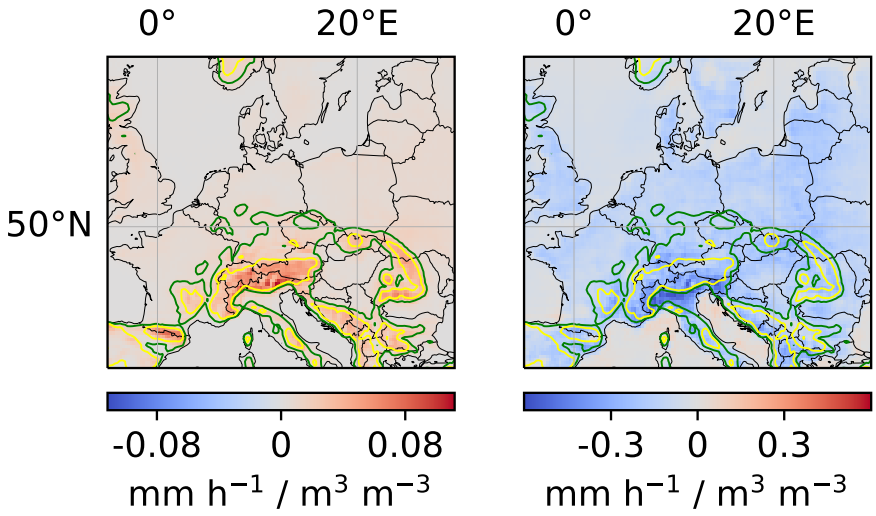

Figure A5. Sum of local and regional soil moisture-convective precipitation and soil moisture-large-scale precipitation couplings. Left: sum of local couplings. Right: sum of regional couplings. See Appendix Fig. A6 for soil moisture-convective precipitation and soil moisture-large-scale precipitation couplings.

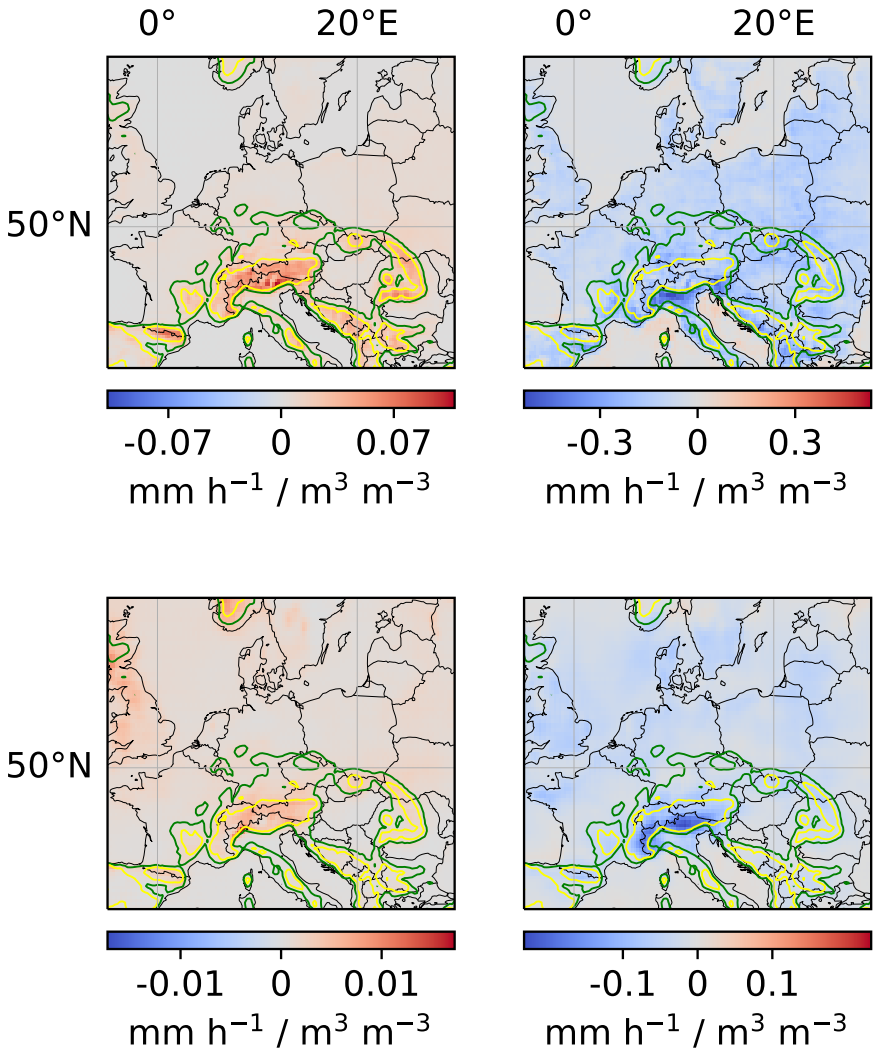

Figure A6. Local and regional soil moisture-convective precipitation and soil moisture-large-scale precipitation couplings. Left column: local couplings. Right column: regional couplings. Upper row: soil moisture-convective precipitation coupling. Lower row: soil moisture-large-scale precipitation coupling.

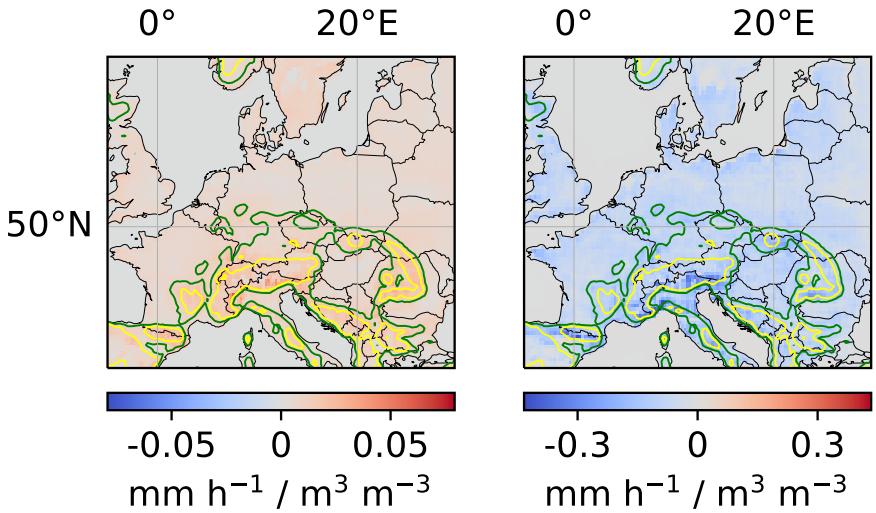

Figure A7. Product of local soil moisture-evaporation and local/ regional evaporation-precipitation couplings. Left: product of local soil moisture-evaporation and local evaporation-precipitation couplings. Right: product of local soil moisture-evaporation and regional evaporation-precipitation couplings. See Appendix Fig. A8 for local soil moisture-evaporation and local and regional evaporation-precipitation couplings.

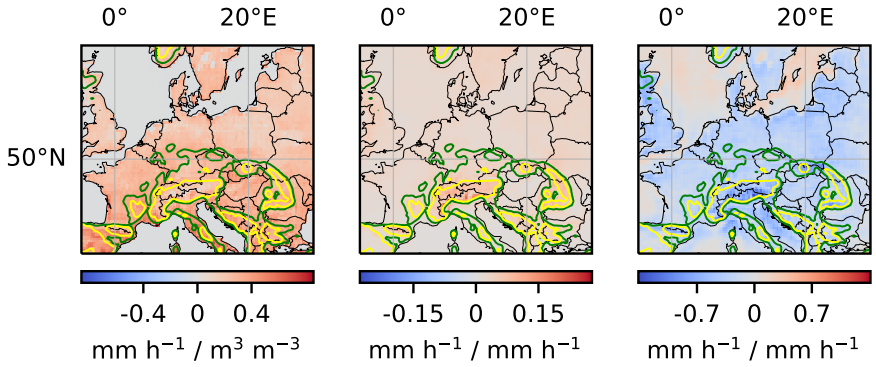

Figure A8. Local soil moisture-evaporation and local and regional evaporation-precipitation couplings. Left: local soil moisture-evaporation coupling. Centre: local evaporation-precipitation coupling. Right: regional evaporation-precipitation coupling.

Author contributions. TT and SK designed the study and analyzed the results with contributions from JG. TT conducted the experiments. TT prepared the manuscript with contributions from SK and JG.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We acknowledge Andreas Hense for valuable discussions on the significance analysis. Further, we gratefully acknowledge the computing time granted through JARA on the supercomputer JURECA at Forschungszentrum Jülich and the Earth System Modelling Project (ESM) for funding this work by providing computing time on the ESM partition of the supercomputer JUWELS at the Jülich Supercomputing Centre (JSC). The work described in this paper received funding from the Helmholtz-RSF Joint Research Group through the project 'European hydro-climate extremes: mechanisms, predictability and impacts', the Initiative and Networking Fund of the Helmholtz Association (HGF) through the project 'Advanced Earth System Modelling Capacity (ESM)', and the Fraunhofer Cluster of Excellence 'Cognitive Internet Technologies'. The content of the paper is the sole responsibility of the author(s) and it does not represent the opinion of the Helmholtz Association, and the Helmholtz Association is not responsible for any use that might be made of the information contained. The ERA5 climate reanalysis dataHersbach et al. (2018) were downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store. The results contain modified Copernicus Climate Change Service information 2021. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

References

Adler, B., Kalthoff, N., and Gantner, L.: Initiation of deep convection caused by land-surface inhomogeneities in West Africa: a modelled case study, Meteorology and Atmospheric Physics, 112, 15–27, https://doi.org/10.1007/s00703-011-0131-2, 2011.

Barnes, E. A., Samarasinghe, S. M., Ebert-Uphoff, I., and Furtado, J. C.: Tropospheric and Stratospheric Causal Pathways Between the MJO and NAO, Journal of Geophysical Research: Atmospheres, 124, 9356–9371, https://doi.org/10.1029/2019jd031024, 2019.

Baur, F., Keil, C., and Craig, G. C.: Soil moisture–precipitation coupling over Central Europe: Interactions between surface anomalies at different scales and the dynamical implication, Quarterly Journal of the Royal Meteorological Society, 144, 2863–2875, https://doi.org/10.1002/qj.3415, 2018.

Dumoulin, V. and Visin, F.: A guide to convolution arithmetic for deep learning, https://arxiv.org/abs/1603.07285, 2016.

Ebert-Uphoff, I. and Deng, Y.: Causal discovery in the geosciences—Using synthetic data to learn how to interpret results, Computers & Geosciences, 99, 50–60, https://doi.org/10.1016/j.cageo.2016.10.008, 2017.

Ebert-Uphoff, I. and Hilburn, K.: Evaluation, Tuning, and Interpretation of Neural Networks for Working with Images in Meteorological Applications, Bulletin of the American Meteorological Society, 101, E2149–E2170, https://doi.org/10.1175/bams-d-20-0097.1, 2020.

Eltahir, E. A. B.: A Soil Moisture–Rainfall Feedback Mechanism: 1. Theory and observations, Water Resources Research, 34, 765–776, https://doi.org/10.1029/97WR03499, 1998.

Findell, K. L. and Eltahir, E. A. B.: Atmospheric Controls on Soil Moisture–Boundary Layer Interactions. Part I: Framework Development, Journal of Hydrometeorology, 4, 552–569, https://doi.org/10.1175/1525-7541(2003)004<0552:acosml>2.0.co;2, 2003a.

Findell, K. L. and Eltahir, E. A. B.: Atmospheric Controls on Soil Moisture–Boundary Layer Interactions. Part II: Feedbacks within the Continental United States, Journal of Hydrometeorology, 4, 570–583, https://doi.org/10.1175/1525-7541(2003)004<0570:acosml>2.0.co;2, 2003b.

Froidevaux, P., Schlemmer, L., Schmidli, J., Langhans, W., and Schär, C.: Influence of the Background Wind on the Local Soil Moisture–Precipitation Feedback, Journal of the Atmospheric Sciences, 71, 782–799, https://doi.org/10.1175/jas-d-13-0180.1, 2014.

Gagne II, D. J., Haupt, S. E., Nychka, D. W., and Thompson, G.: Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms, Monthly Weather Review, 147, 2827–2845, https://doi.org/10.1175/mwr-d-18-0316.1, 2019.

Gentine, P., Holtslag, A. A. M., D'Andrea, F., and Ek, M.: Surface and Atmospheric Controls on the Onset of Moist Convection over Land, Journal of Hydrometeorology, 14, 1443–1462, https://doi.org/10.1175/jhm-d-12-0137.1, 2013.

Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M., and Kagal, L.: Explaining Explanations: An Overview of Interpretability of Machine Learning, in: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80–89, IEEE, https://doi.org/10.1109/dsaa.2018.00018, 2018.

Green, J. K., Konings, A. G., Alemohammad, S. H., Berry, J., Entekhabi, D., Kolassa, J., Lee, J.-E., and Gentine, P.: Regionally strong feedbacks between the atmosphere and terrestrial biosphere, Nat Geosci, 10, 410–414, https://doi.org/10.1038/ngeo2957, 2017.

Green, J. K., Seneviratne, S. I., Berg, A. M., Findell, K. L., Hagemann, S., Lawrence, D. M., and Gentine, P.: Large influence of soil moisture on long-term terrestrial carbon uptake, Nature, 565, 476–479, https://doi.org/10.1038/s41586-018-0848-x, 2019.

Guillod, B. P., Orlowsky, B., Miralles, D. G., Teuling, A. J., and Seneviratne, S. I.: Reconciling spatial and temporal soil moisture effects on afternoon rainfall, Nat Commun, 6, https://doi.org/10.1038/ncomms7443, 2015.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H.: A Survey of Learning Causality with Data, ACM Computing Surveys, 53, 1–37, https://doi.org/10.1145/3397269, 2021.

Ham, Y., Kim, J., and Luo, J.: Deep learning for multi-year ENSO forecasts, Nature, 573, 568–572, https://doi.org/10.1038/s41586-019-1559-7, 2019.

Hartick, C., Furusho-Percot, C., Goergen, K., and Kollet, S.: An Interannual Probabilistic Assessment of Subsurface Water Storage Over Europe Using a Fully Coupled Terrestrial Model, Water Resources Research, 57, https://doi.org/10.1029/2020wr027828, 2021.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on 18-06-2021), https://doi.org/http://dx.doi.org/10.24381/cds.adbb2d47, 2018.

Hesterberg, T.: What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum, https://arxiv.org/abs/1411.5279, 2014.

Holgate, C. M., Dijk, A. I. J. M. V., Evans, J. P., and Pitman, A. J.: The Importance of the One-Dimensional Assumption in Soil Moisture - Rainfall Depth Correlation at Varying Spatial Scales, Journal of Geophysical Research: Atmospheres, 124, 2964–2975, https://doi.org/10.1029/2018jd029762, 2019.

Humphrey, V., Berg, A., Ciais, P., Gentine, P., Jung, M., Reichstein, M., Seneviratne, S. I., and Frankenberg, C.: Soil moisture–atmosphere feedback dominates land carbon uptake variability, Nature, 592, 65–69, https://doi.org/10.1038/s41586-021-03325-5, 2021.

Imamovic, A., Schlemmer, L., and Schär, C.: Collective impacts of orography and soil moisture on the soil moisture-precipitation feedback, Geophysical Research Letters, 44, 11,682–11,691, https://doi.org/10.1002/2017GL075657, 2017.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, https://arxiv.org/abs/1412.6980, 2017.

Koster, R. D.: Regions of Strong Coupling Between Soil Moisture and Precipitation, Science, 305, 1138–1140, https://doi.org/10.1126/science.1100217, 2004.

LeCun, Y., Bengio, Y., and Hinton, G.: Deep learning, Nature, 521, 436–444, https://doi.org/10.1038/nature14539, 2015.

Leutwyler, D., Imamovic, A., and Schär, C.: The Continental-Scale Soil-Moisture Precipitation Feedback in Europe with Parameterized and Explicit Convection, Journal of Climate, 34, 1–56, https://doi.org/10.1175/jcli-d-20-0415.1, 2021.

Massmann, A., Gentine, P., and Runge, J.: Causal inference for process understanding in Earth sciences, https://arxiv.org/abs/2105.00912, 2021.

McGovern, A., Lagerquist, R., Gagne II, D. J., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., and Smith, T.: Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning, Bulletin of the American Meteorological Society, 100, 2175–2199, https://doi.org/10.1175/bams-d-18-0195.1, 2019.

Miller, J. W., Goodman, R., and Smyth, P.: On loss functions which minimize to conditional expected values and posterior probabilities, IEEE Transactions on Information Theory, 39, 1404–1408, https://doi.org/10.1109/18.243457, 1993.

Molnar, C.: Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/, 2019.

Montavon, G., Samek, W., and Müller, K.: Methods for interpreting and understanding deep neural networks, Digital Signal Processing, 73, 1–15, https://doi.org/10.1016/j.dsp.2017.10.011, 2018.

Padarian, J., McBratney, A. B., and Minasny, B.: Game theory interpretation of digital soil mapping convolutional neural networks, SOIL, 6, 389–397, https://doi.org/10.5194/soil-6-389-2020, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems 32, edited by Wallach, H.,

Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., pp. 8026–8037, Curran Associates, Inc., http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf, 2019.

Pearl, J.: Causal inference in statistics: An overview, Statistics Surveys, 3, https://doi.org/10.1214/09-ss057, 2009.

Peters, J., Bühlmann, P., and Meinshausen, N.: Causal inference by using invariant prediction: identification and confidence intervals, J. R. Stat. Soc.: Series B (Statistical Methodology), 78, 947–1012, https://doi.org/10.1111/rssb.12167, 2016.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Ronneberger, O., Fischer, P., and Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, edited by Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., pp. 234–241, Springer International Publishing, Cham, https://arxiv.org/abs/1505.04597, 2015.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J.: Explainable Machine Learning for Scientific Insights and Discoveries, IEEE Access, 8, 42 200–42 216, https://doi.org/10.1109/ACCESS.2020.2976199, 2020.

Runge, J.: Causal network reconstruction from time series: From theoretical assumptions to practical estimation, Chaos: An Interdisciplinary Journal of Nonlinear Science, 28, 075 310, https://doi.org/10.1063/1.5025050, 2018.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., van Nes, E. H., Peters, J., Quax, R., Reichstein, M., Scheffer, M., Schölkopf, B., Spirtes, P., Sugihara, G., Sun, J., Zhang, K., and Zscheischler, J.: Inferring causation from time series in Earth system sciences, Nat Commun, 10, https://doi.org/10.1038/s41467-019-10105-3, 2019.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K. R.: Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications, Proceedings of the IEEE, 109, 247–278, https://doi.org/10.1109/JPROC.2021.3060483, 2021.

Santanello, J. A., Dirmeyer, P. A., Ferguson, C. R., Findell, K. L., Tawfik, A. B., Berg, A., Ek, M., Gentine, P., Guillod, B. P., van Heerwaarden, C., Roundy, J., and Wulfmeyer, V.: Land–Atmosphere Interactions: The LoCo Perspective, Bulletin of the American Meteorological Society, 99, 1253–1272, https://doi.org/10.1175/bams-d-17-0001.1, 2018.

Schumacher, D. L., Keune, J., van Heerwaarden, C. C., de Arellano, J. V.-G., Teuling, A. J., and Miralles, D. G.: Amplification of mega-heatwaves through heat torrents fuelled by upwind drought, Nature Geoscience, 12, 712–717, https://doi.org/10.1038/s41561-019-0431-6, 2019.

Schwingshackl, C., Hirschi, M., and Seneviratne, S. I.: Quantifying Spatiotemporal Variations of Soil Moisture Control on Surface Energy Balance and Near-Surface Air Temperature, Journal of Climate, 30, 7105–7124, https://doi.org/10.1175/jcli-d-16-0727.1, 2017.

Seneviratne, S. I., Lüthi, D., Litschi, M., and Schär, C.: Land–atmosphere coupling and climate change in Europe, Nature, 443, 205–209, https://doi.org/10.1038/nature05095, 2006.

Seneviratne, S. I., Corti, T., Davin, E. L., Hirschi, M., Jaeger, E. B., Lehner, I., Orlowsky, B., and Teuling, A. J.: Investigating soil moisture–climate interactions in a changing climate: A review, Earth-Science Reviews, 99, 125–161, https://doi.org/10.1016/j.earscirev.2010.02.004, 2010.

Shpitser, I., VanderWeele, T., and Robins, J. M.: On the Validity of Covariate Adjustment for Estimating Causal Effects, in: Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI'10, p. 527–536, AUAI Press, Arlington, Virginia, USA, 2010.

Taylor, C. M.: Detecting soil moisture impacts on convective initiation in Europe, Geophysical Research Letters, 42, 4631–4638, https://doi.org/10.1002/2015gl064030, 2015.

Taylor, C. M., Gounou, A., Guichard, F., Harris, P. P., Ellis, R. J., Couvreux, F., and Kauwe, M. D.: Frequency of Sahelian storm initiation enhanced over mesoscale soil-moisture patterns, Nature Geoscience, 4, 430–433, https://doi.org/10.1038/ngeo1173, 2011.

Tesch, T., Kollet, S., and Garcke, J.: Variant Approach for Identifying Spurious Relations That Deep Learning Models Learn, Frontiers in Water, 3, 114, https://doi.org/10.3389/frwa.2021.745563, 2021.

Tietz, M., Fan, T. J., Nouri, D., Bossan, B., and skorch Developers: skorch: A scikit-learn compatible neural network library that wraps PyTorch, https://skorch.readthedocs.io/en/stable/, 2017.

Toms, B. A., Barnes, E. A., and Ebert-Uphoff, I.: Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability, Journal of Advances in Modeling Earth Systems, 12, e2019MS002 002, https://doi.org/10.1029/2019ms002002, 2020.

Tuttle, S. and Salvucci, G.: Empirical evidence of contrasting soil moisture–precipitation feedbacks across the United States, Science, 352, 825–828, https://doi.org/10.1126/science.aaa7185, 2016.

Tuttle, S. E. and Salvucci, G. D.: Confounding factors in determining causal soil moisture-precipitation feedback, Water Resources Research, 53, 5531–5544, https://doi.org/10.1002/2016wr019869, 2017.

Welty, J. and Zeng, X.: Does Soil Moisture Affect Warm Season Precipitation Over the Southern Great Plains?, Geophysical Research Letters, 45, 7866–7873, https://doi.org/10.1029/2018gl078598, 2018.

Witte, J., Henckel, L., Maathuis, M. H., and Didelez, V.: On Efficient Adjustment in Causal Graphs, Journal of Machine Learning Research, 21, 1–45, https://doi.org/10.48550/arXiv.2002.06825, 2020.

Zhang, Q. and Zhu, S.: Visual interpretability for deep learning: a survey, Frontiers Inf Technol Electronic Eng, 19, 27–39, https://doi.org/10.1631/fitee.1700808, 2018.