Causal deep learning models for studying the Earth system~~: soil moisture-precipitation coupling in ERA5 data across Europe~~

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. ~~The Earth system~~ 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 ~~are still~~ remain poorly understood. Current approaches for studying relations in the Earth system ~~may be broadly divided into approaches based~~ rely either on numerical simulations ~~and~~ or statistical approaches. However, there are several inherent limitations to ~~current approachesthat are, for example,~~ existing approaches, including high computational costs, ~~reliance on the correct representation of relations~~ uncertainties in numerical models, strong assumptions ~~related to~~ 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. ~~The methodology combines the~~ On the one hand, we employ the recent idea of training and analyzing DL models to gain new scientific insights ~~in the~~ into relations between input and target variables~~with a theorem from causality research. This theorem states~~ . On the other hand, we use that a statistical model ~~may learn the causal impact~~ 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 ~~is a dynamical system featuring~~ comprises many complex processes and non-linear relations between ~~different Earth system variables. Despite many years of research, sophisticated numerical models and a plethora of observational data, many of these processes and relations are still poorly understood. Consider~~ 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~~ , 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 ~~(e.g. Seneviratne et al., 2010; Santanello et al., 2018)~~ (Seneviratne et al., 2010; Santanello et al., 2018). However, because there are several concurring pathways of soil moisture-

precipitation coupling~~(see upper panel of Fig. ??)~~, it remains an open question whether an increase in soil moisture leads to an increase or decrease in precipitation. Answering this question ~~is important, because a better understanding of soil moisture-precipitation coupling might improve precipitation predictions~~ might lead to improved precipitation predictions with numerical models.

~~Certainly, there are many approaches~~ Approaches for studying relations in the Earth system ~~. These approaches~~ 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). ~~However, current~~ Both classes of approaches have several inherent limitations. ~~On the one hand, approaches~~ 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 ~~parameterizations, such that using these numerical~~ simplified parameterizations, thus, using these models to study soil moisture-precipitation coupling ~~may not be optimal~~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 ~~often bring their ownlimitations~~have strong limitations on their own, for example ~~strong assumptions like~~ due to assumptions on linearity or locality of ~~the~~ considered relations and negligence of the ~~discrepancy~~ 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 ~~deep learning (e.g. Zhang and Zhu, 2018; Montavon et al., 2018; Gilpin et al., 2018; Molnar, 2019; Samek et al., 2021)~~ 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 ~~was~~ 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 ~~power and flexibility~~ use of DL models allows to overcome common assumptions in other statistical approaches like linearity or locality. So far, however, the ~~discrepancy~~ difference between causality and correlation has been neglected in ~~the~~ studies using this approach. Indeed, DL models might learn ~~all kinds of~~ various (spurious) correlations between input and target variables, while researchers striving for new scientific insights are most interested in ~~the causal ones. Thus, we propose to extend~~ causal relations.

Therefore, in this work, we propose extending the approach by combining it with a ~~theorem~~ result from causality research ~~that states that~~ stating that a statistical model may learn the causal ~~impact~~ effect of an input variable on a target variable if suitable additional input variables are included ~~(Pearl, 2009).~~

~~Although there exist several 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., 20~~ ~~, most of them focus on discovering causal links and estimating the structure of unknown causal graphs. The formal statement from (Pearl, 2009) that we apply in the proposed methodology~~ (Pearl, 2009; Shpitser et al., 2010). In the geosciences, this result has only recently received attention in the work of (Massmann et al., 2021)~~and has not yet been~~ . In this work, it is combined with the methodology of training and analyzing DL models to gain new scientific insights ~~. Harnessing the great~~

power and flexibility of DL models, this combination can yield new scientific insights into complex, nonlinear and non-local mechanisms in the Earth system.

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., 20 . 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. While we believe that our results on soil moisture-precipitation coupling may contribute to a better understanding of this coupling, in this article, we focus on demonstrating the methodology. An extensive discussion of our results on soil moisture-precipitation coupling in terms of physical processes (e.g. Seneviratne et al., 2010; Santanello et al., 2018) and a comparison with results from other studies (e.g. Seneviratne et al., 2010; Taylor et al., 2012; Guillod et al., 2015; Tuttle and Salvucci, 2016; Imamovic et al., 2017) are postponed to a second paperOther 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).

The This manuscript is structured such that as follows: Sect. 2 introduces the required background on causality research and details the proposed methodology, . Sect. 3 presents the application of the methodology to the example of to soil moisture-precipitation coupling , and and provides a comparison to other approaches. Finally, Sect. 4 presents several further contains several additional analyses to assess the statistical significance and correctness of results obtained with the proposed methodology. Finally, Sect. 3.1 compares soil moisture-precipitation coupling obtained from

2 Methodology

To introduce the proposed methodologywith soil moisture-precipitation coupling obtained from a simple linear correlation analysis. , 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.

3 Methodology

Figure ?? provides a conceptual overview of the proposed methodology. Given a complex relation between two variables, for example soil moisture-precipitation coupling, we train a *causal* DL model to predict one variable given the other, and perform a sensitivity analysis of the trained model to analyze how the target variablechanges when the respective input variable is changed.

In this section, we introduce the required background on causality research and detail the proposed methodology. In particular, Sect. 2.1 clarifies what is meant by *causal impact* and presents the formal statement from (Pearl, 2009) that a statistical

model may learn the causal impact of an input variable $\boldsymbol{X} \in \mathbb{R}^d$ on a target variable $\boldsymbol{Y} \in \mathbb{R}^n$ if suitable additional input variables $\boldsymbol{C_i} \in \mathbb{R}^{d_i}, i = 1, \ldots, k$, are chosen. Subsequently, Sect. 2.2 details the proposed methodology, i.e. the training of a causal DL model and the sensitivity analysis of the trained model.

Schematic of the methodology in general (text) and in the example of soil moisture-precipitation coupling (figures). The upper figure depicts different effects of soil moisture increases and the corresponding impact on precipitation. The lower left figure depicts the DL model considered in our example, and the lower right figure shows an exemplary result of the sensitivity analysis.

2.1 Causal background

2.1 Background on causality

The *causal impact* of 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) expected response of $\boldsymbol{Y}$ to intervening into the considered system (e.g. the Earth system) and changing the value of $\boldsymbol{X}$. To better understand soil moisture-precipitation coupling, one might for example be interested in the expected response of precipitation to intervening into the Earth system and increasing or decreasing soil moisture across Europe. In order to determine the causal impactof variable $\boldsymbol{X}$ on $\boldsymbol{Y}$ determine this impact, one has to determine the expected value of $\boldsymbol{Y}$ given that one intervened into the system and set sets $\boldsymbol{X}$ to some arbitrary value $\boldsymbol{x}$. In the framework of Structural Causal Models (SCMs) introduced below, this 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 $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})]$. Note that, in general, it holds $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})] \neq \mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}]$, which will be discussed below. Note further that here and in the following, small letters $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{c_i}$ refer to particular values of the random variables $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{C_i}$, respectively.

In some cases, the value $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})]$ can be computed by actually intervening into the considered systemdetermined experimentally by setting $\boldsymbol{X}$ to $\boldsymbol{x}$ while monitoring $\boldsymbol{Y}$. For example, when the considered system is a numerical model of the Earth system, one might compute $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})]$ by performing several simulations with randomly perturbed soil moisture. However, when we do not want to rely on numerical simulations (e.g. due to computational constraints), but directly consider the Earth system, performing the required interventions for computing $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})]$ may not be possible, e.g. it is not possible to intervene in the Earth system and randomly increase or decrease soil moisture on a large scale. Nevertheless, even when it is not possible to intervene into the system, it is often still possible to determine $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})]$, i.e. the expected value of $\boldsymbol{Y}$ if we *would* intervene in the system and set in Earth System Modeling (ESM), one may be able to set $\boldsymbol{X}$ to $\boldsymbol{x}$, and thus the causal impact of $\boldsymbol{X}$ on $\boldsymbol{Y}$ $\boldsymbol{x}$ 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.

~~In the following, we take a brief look at the framework of Structural Causal Models (SCMs), which gives~~ The framework of SCMs (Pearl, 2009) provides a deeper understanding of the notion $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})]$, and ~~how to determine $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})]$ in the case that we cannot intervene in the system. Note that we simplify some parts to focus on the aspects that are important for the proposed methodology.~~ 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 ~~to the framework~~ we refer to (Pearl, 2009). ~~Another~~ An introduction to the framework in the context of ~~Earth sciences~~ geosciences is given in (Massmann et al., 2021).

~~Underlying~~

2.1.1 Structural Causal Models

In the framework of SCMs~~is the concept of causal graphs~~, 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 (DAG) that encodes our assumptions about the causal dependencies of a system (see left panel of Fig. 1 for an example and terminology). The nodes of the graph represent variablesof the system , while adirected edge from some node $\boldsymbol{A}$ to another node $\boldsymbol{B}$ represents a *direct* causal impact of variable $\boldsymbol{A}$ on variable $\boldsymbol{B}$.~~ 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.

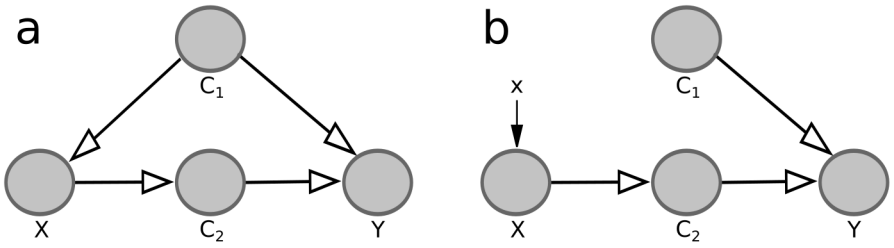

Figure 1. ~~**Example for a causal graph (left) and corresponding causal graph for intervening into the system and setting variable X to some value x_0 (right).** A causal graph is a Directed Acyclic Graph (DAG) that encodes our assumptions about the causal dependencies of a system.~~ **Example for a 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~~and represent variables of the system~~, while the arrows are referred to as ~~(~~directed ~~)~~ edges.~~A directed edge from some node $\boldsymbol{A}$ to another node $\boldsymbol{B}$ represents a *direct* causal impact of variable $\boldsymbol{A}$ on variable $\boldsymbol{B}$.~~

Formally, ~~it is assumed that the value of some variable $\boldsymbol{A}$~~ a variable in the causal graph is determined by a ~~(deterministic) function f_A~~function f, whose inputs are ~~the parents of node $\boldsymbol{A}$, i.e. all nodes with an edge pointing to $\boldsymbol{A}$, and an independent variable $\boldsymbol{U_A}$~~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 ~~the left panel of~~ Fig. 1~~, it is assumed that~~ a, the four variables are determined by four

functions $f_{C_1}, f_{C_2}, f_X, f_Y$ ~~, such that:~~

$$\boldsymbol{c_1} = f_{C_1}(\boldsymbol{u_{C_1}})$$

$$\boldsymbol{x} = f_X(\boldsymbol{c_1}, \boldsymbol{u_X})$$

$$\boldsymbol{c_2} = f_{C_2}(\boldsymbol{x}, \boldsymbol{u_{C_2}})$$

$$\boldsymbol{y} = f_Y(\boldsymbol{c_1}, \boldsymbol{c_2}, \boldsymbol{u_Y}),$$

$$\vdots$$

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

~~where the $\boldsymbol{U_{C_1}}, \boldsymbol{U_{C_2}}, \boldsymbol{U_X}, \boldsymbol{U_Y}$ represent other variables of the system (or noise) and are assumed to be jointly independent random variables.~~

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 ~~joint~~ 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.

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

~~Above, we stated that the causal impact of some variable $\boldsymbol{X} \in \mathbb{R}^d$ on another variable $\boldsymbol{Y} \in \mathbb{R}^n$ is the (expected) response~~ As stated above, to determine the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$, one has to determine the expected value of $\boldsymbol{Y}$ ~~to intervening into the considered system and changing the value~~ 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, ~~the intervention into the system~~ this corresponds to removing all edges in the causal graph pointing to $\boldsymbol{X}$, and modifying the structural equation for $\boldsymbol{X}$ ~~. To study the causal impact of variable~~ accordingly. For example, when studying the causal effect of $\boldsymbol{X}$ on ~~variable~~ $\boldsymbol{Y}$ in ~~the left panel of~~ Fig. 1~~, for example, we might intervene in the system and set $\boldsymbol{X}$ to some constant $\boldsymbol{x_0}$. The modified system would be~~ a, the modified system is described by the causal graph in ~~the right panel of~~ Fig. 1~~, together with the structural equations~~

$$\boldsymbol{c_1} = f_{C_1}(\boldsymbol{u_{C_1}})$$

$$\boldsymbol{x} = \boldsymbol{x_0}$$

$$\boldsymbol{c_2} = f_{C_2}(\boldsymbol{x}, \boldsymbol{u_{C_2}})$$

$$\boldsymbol{y} = f_Y(\boldsymbol{c_1}, \boldsymbol{c_2}, \boldsymbol{u_Y}).$$

b 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 ~~, denoted $\mathbb{P}(\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{Y} | do(\boldsymbol{X} = \boldsymbol{x_0}))$, and corresponding expected value $\mathbb{E}[\boldsymbol{Y} | do(\boldsymbol{X} = \boldsymbol{x_0})]$ of $\boldsymbol{Y}$ given that one intervened into the system and set~~ $\mathbb{P}(\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{Y} | do(\boldsymbol{X} = \boldsymbol{x}))$, 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 $\boldsymbol{X}$ ~~to $\boldsymbol{x_0}$. If we could observe this modified system (e.g. by actually intervening into the Earth system)~~on $\boldsymbol{Y}$ and differs from the expected value for the original system, ~~we could study the causal impact of $\mathbb{E}[\boldsymbol{Y} | \boldsymbol{X} = \boldsymbol{x}]$. 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 $\boldsymbol{Y}$ ~~by analyzing quantities such as~~

$$\mathbb{E}[\boldsymbol{Y} | do(\boldsymbol{X} = \boldsymbol{x_0})] - \mathbb{E}[\boldsymbol{Y} | do(\boldsymbol{X} = \boldsymbol{x_1})].$$

~~However, if we cannot intervene in the system, we can only observe the original systemand the distribution $\mathbb{P}(\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{Y} | \boldsymbol{X} = \boldsymbol{x_0})$. To illustrate the difference between $\mathbb{P}(\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{Y} | do(\boldsymbol{X} = \boldsymbol{x_0}))$ and $\mathbb{P}(\boldsymbol{C_1}, \boldsymbol{C_2}, \boldsymbol{Y} | \boldsymbol{X} = \boldsymbol{x_0})$, note that in the latter case, the value of~~ (via its impact on $\boldsymbol{C_2}$). However, in the original system, knowing $\boldsymbol{X}$ allows to draw ~~conclusions about the value of~~ additional conclusions about $\boldsymbol{C_1}$~~, because~~. 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 ~~the~~ $\boldsymbol{X}$, 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 $\boldsymbol{X}$ ~~. That again~~ 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 ~~the value of~~ $\boldsymbol{Y}$ ~~,~~ because $\boldsymbol{C_1}$ also affects ~~the value of~~ $\boldsymbol{Y}$ ~~(~~. Summarizing, due to the confounding influence of $\boldsymbol{C_1}$ ~~is a *confounder*). In the former case, on~~ , 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 ~~other hand, we intervene in the system and~~ cases considered in the proposed methodology, it is impossible or undesirable to experimentally set $\boldsymbol{X}$ to ~~some arbitrary value $\boldsymbol{x_0}$. In this case, the value of $\boldsymbol{X}$ does not allow to draw any conclusions about the value of $\boldsymbol{C_1}$~~$\boldsymbol{x}$. 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 $\mathbb{E}[\boldsymbol{Y} | do(\boldsymbol{X} = \boldsymbol{x})]$.

~~To study the causal impact of some variable~~

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 $\boldsymbol{X}$ ~~on another variable and~~ $\boldsymbol{Y}$ ~~when we cannot intervene in the system, we need to bridge the gap between the distributions $\mathbb{P}(\boldsymbol{C_1},\boldsymbol{C_2},\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x_0}))$ and $\mathbb{P}(\boldsymbol{C_1},\boldsymbol{C_2},\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x_0})$. To do so, the proposed methodology relies on the following theorem from causality research (Pearl, 2009):~~

~~**Theorem 1:** For multi-valued variables $\boldsymbol{X}\in\mathbb{R}^d,\boldsymbol{Y}\in\mathbb{R}^n$, finding a sufficient set S of multi-valued variables $\boldsymbol{C_i}\in\mathbb{R}^{d_i}, i=1,\dots,k,$ permits us to write~~

$$\mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|do(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C_i}=\boldsymbol{c_i}\}_{i=1}^k)=\mathbb{P}(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}=\boldsymbol{x},\{\boldsymbol{C_i}=\boldsymbol{c_i}\}_{i=1}^k).$$

~~Note that this implies~~

$$\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C_i}=\boldsymbol{c_i}\}_{i=1}^k]=\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\{\boldsymbol{C_i}=\boldsymbol{c_i}\}_{i=1}^k].$$

~~A sufficient set is defined as follows:~~

~~**Definition 1 (Sufficient set):** In the context of Theorem 1, a set S of multi-valued variables $\boldsymbol{C_i}\in\mathbb{R}^{d_i}, i=1,\dots,k,$ is sufficient if : No element of S is a descendant of~~ . Indeed, in the example from Fig. 1, we showed that original and post-intervention expected values differ because, in the original system, knowing $\boldsymbol{X}$ ~~. The elements of S block all paths between~~ 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 ~~$\boldsymbol{Y}$ that contain an edge pointing to~~ $\boldsymbol{C_1}$. 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 $\boldsymbol{x}$. ~~Here, a descendant of $\boldsymbol{X}$ is any variable $\boldsymbol{D}$ for which there exists a directed path $\boldsymbol{X}\to\dots\to\boldsymbol{D}$~~

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 $\boldsymbol{C_\ell}\in\mathbb{R}^{d_\ell}, \ell=1,\dots,k,$ 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 ~~$\boldsymbol{D}$ in the~~ $\boldsymbol{Y}$ in the original causal graph. ~~In the left panel of Fig. 1 for example, the variables $\boldsymbol{C_2}$ and $\boldsymbol{Y}$ are descendants of~~

2. No $\{\boldsymbol{C_\ell}\}_{\ell=1}^k$ lies on a causal path from $\boldsymbol{X}$ ~~, while $\boldsymbol{C_1}$ is not. In the second condition~~to $\boldsymbol{Y}$.

Here, a path is any ~~sequence "node-edge-node-edge-...-edge-node", where the edgesdo not necessarily all point in the same direction~~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 ~~p is blocked~~ is *blocked* by a set ~~S~~ $S = \{\boldsymbol{C_\ell}\}_{\ell=1}^k$ of nodes if either (i) ~~p~~ 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) ~~p~~ the path contains at least one collision node, i.e. a node ~~on the path~~ with both adjacent edges pointing towards the node ($\ldots \rightarrow \boldsymbol{C} \leftarrow \ldots$), which is outside S and has no descendants in S ~~. In~~ (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 ~~soil moisture-precipitation coupling, we give some intuition on these conditions. For further details, we refer to (Pearl, 2009). Note that the parents~~ $\boldsymbol{C_1}$ 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 $\boldsymbol{X}$ ~~, i.e.~~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 $\boldsymbol{C_\ell} \in \mathbb{R}^{d_\ell}, \ell = 1, \ldots, k$ 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. ~~all variables which have a direct impact on $\boldsymbol{X}$ (in~~ on the structural equations. Note that the parents of $\boldsymbol{X}$ in the causal graph always fulfil the adjustment criteria. In the ~~causal graphrepresented by an edge pointing from the respective variable to $\boldsymbol{X}$), always form a sufficient set.~~ 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

~~This section details the proposed methodology. Figure **??** provides a conceptual overview 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 ~~$\boldsymbol{C_i} \in \mathbb{R}^{d_i}, i = 1, \ldots, k,$ and~~ $\boldsymbol{C_\ell} \in \mathbb{R}^{d_\ell}, \ell = 1, \ldots, k$. In a second step, we perform a sensitivity analysis of the trained model ~~. The sensitivity analysis answers the question~~ to analyze how $\boldsymbol{Y}$ ~~changes when~~ would change if we changed $\boldsymbol{X}$, i.e. to determine the causal effect of $\boldsymbol{X}$ ~~is changed . Section 2.2.1 details the procedure of training a causal deep learning model, while the sensitivity analysis is detailed in Sect. 2.2.2.~~ on $\boldsymbol{Y}$.

2.2.1 Training a *causal* DL model ~~to predict one variable given the other~~

DL models (LeCun et al., 2015; Reichstein et al., 2019) learn statistical associations between their input and target variables. By training a ~~*causal*~~ causal DL model, we mean that we train a DL model ~~which approximates the map~~

$$(\boldsymbol{x}, \{\boldsymbol{c_i}\}_{i=1}^k) \rightarrow \mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \{\boldsymbol{C_i} = \boldsymbol{c_i}\}_{i=1}^k],$$

~~where $\boldsymbol{Y} \in \mathbb{R}^n, \boldsymbol{X} \in \mathbb{R}^d$ and $\boldsymbol{C_i} \in \mathbb{R}^{d_i}, i = 1, \ldots, k$ (see Sect. 2.1 for an explanation of the notion $do(\boldsymbol{X} = \boldsymbol{x})$). In Sect. 2.2.2, we will use this model to determine the causal impact of $\boldsymbol{X}$ on $\boldsymbol{Y}$.~~

~~To achieve that the DL~~ 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 ~~from Eq. 5~~

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

To obtain a causal DL model, the loss function, ~~DL model~~ model architecture and additional input variables ~~$\boldsymbol{C_i}, i = 1, \ldots, k,$~~ $\{\boldsymbol{C_\ell}\}_{\ell=1}^k$ 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

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

~~A loss function fulfilling this requirement is , for example,~~ An example for such a loss function is the expected mean squared error,

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

~~where $\hat{\boldsymbol{Y}}$ is the model's prediction of $\boldsymbol{Y}$ (Miller et al., 1993). In addition to such a loss function,~~ which maps a function $\boldsymbol{m} : (\boldsymbol{X}, \{\boldsymbol{C_\ell}\}_{\ell=1}^k) \rightarrow \mathbb{R}^n$, 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 ~~$\{\boldsymbol{C_i}\}_{i=1}^k$ that form a sufficient set S (see~~ $\{\boldsymbol{C_\ell}\}_{\ell=1}^k$ that fulfil the adjustment criteria from Sect. ~~2.1). From Theorem 1, we know that this implies $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}, \{\boldsymbol{C_i} = \boldsymbol{c_i}\}_{i=1}^k] = \mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \{\boldsymbol{C_i} = \boldsymbol{c_i}\}_{i=1}^k]$.Note that choosing~~ 2.1.2, such that the maps from Eq. 5 and Eq. 6 become identical. The choice of additional input variables ~~$\{\boldsymbol{C_i}\}_{i=1}^k$ that perfectly fulfill the requirements of a sufficient set in Definition 1 is rarely possible in practice. Nevertheless, in many cases, it might be enough to approximately fulfill the requirements and perform further analyses to assess the correctness of obtained results. We discuss such further analyses in~~ 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. ~~4.~~ 2.1.2), which can be obtained from the proposed methodology.

~~In summary, by choosing a suitable loss function, DL model and additional input variables, we obtain a *causal* DL model, i.e. a DL model that approximates the map from Eq. 5.~~

2.2.2 ~~Performing a sensitivity~~ Sensitivity analysis of the trained model

To determine the causal ~~impact~~ 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.

$$s_{i_1 i_2 ij}(\boldsymbol{x}, \{\boldsymbol{c_\ell}\}_{\ell=1}^k) = \frac{\partial \mathbb{E}[\boldsymbol{Y}_{i_1} | do(\boldsymbol{X} = \boldsymbol{x}), \{\boldsymbol{C_i} = \boldsymbol{c_i}\}_{i=1}^k]}{\partial \boldsymbol{X}_{i_2}} \frac{\partial \mathbb{E}[\boldsymbol{Y}_i | do(\boldsymbol{X} = \boldsymbol{x}), \{\boldsymbol{C_\ell} = \boldsymbol{c_\ell}\}_{\ell=1}^k]}{\partial \boldsymbol{X}_j}, \tag{8}$$

where ~~$i_1 \in \{1, \ldots, n\}$, $i_2 \in \{1, \ldots, d\}$~~ $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, d\}$. These partial derivatives ~~answer how $\boldsymbol{Y}_{i_1}$ changes if we intervened in the considered system and slightly changed the value of $\boldsymbol{X}_{i_2}$. In applications using linear regression models to approximate $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \{\boldsymbol{C_i} = \boldsymbol{c_i}\}_{i=1}^k]$, $s_{i_1 i_2}$ is approximated by the i_1-th linear regression coefficient of $\boldsymbol{X}_{i_2}$ (see, e.g. Pearl, 2009). In our case, however, we have a differentiable DL model that approximates $\mathbb{E}[\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \{\boldsymbol{C_i} = \boldsymbol{c_i}\}_{i=1}^k]$. Accordingly, we take the corresponding partial derivative~~ indicate how $\boldsymbol{Y}_j$ 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 ~~to approximate $s_{i_1 i_2}$.~~ 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_1, i_2)~~ (i, j), for example to analyze the impact of a change in ~~$\boldsymbol{X}_{i_2}$~~ $\boldsymbol{X}_j$ on the sum ~~$\sum_{j=1}^n \boldsymbol{Y}_j$. In the example of~~ $\sum_{i=1}^n \boldsymbol{Y}_i$. When studying soil moisture-precipitation coupling, ~~for instance,~~ we combine different partial derivatives to study the local and regional impact of soil moisture changes on precipitation (see Sect. 3.4).

~~To answer the question, how $\boldsymbol{Y}$ changes on average if we intervened into the system and slightly changed~~ In theory, the proposed methodology identifies the causal effect of $\boldsymbol{X}$ ~~, we consider the expected values of the above partial derivatives with respect to the joint distribution of $\boldsymbol{X}$ and $\{\boldsymbol{C_i}\}_{i=1}^k$, i.e.~~

$$\overline{s_{i_1 i_2}} = \mathbb{E}_{\boldsymbol{x}, \{\boldsymbol{c_i}\}_{i=1}^k}[s_{i_1 i_2}] = \mathbb{E}_{\boldsymbol{x}, \{\boldsymbol{c_i}\}_{i=1}^k}\left[\frac{\partial \mathbb{E}[\boldsymbol{Y}_{i_1} | do(\boldsymbol{X} = \boldsymbol{x}), \{\boldsymbol{C_i} = \boldsymbol{c_i}\}_{i=1}^k]}{\partial \boldsymbol{X}_{i_2}}\right].$$

~~We approximate this quantity by averaging the partial derivatives $s_{i_1 i_2}$ over a large number of observed tuples $(\boldsymbol{x}, \{\boldsymbol{c_i}\}_{i=1}^k)$. For instance, when studying soil moisture-precipitation coupling, we average~~ 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 ~~of the trained DL model over all samples from the test set.~~ $q_{ij}(\boldsymbol{x}, \{\boldsymbol{c}_\ell\}_{\ell=1}^k)$ 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 ~~example~~to soil moisture-precipitation coupling

~~This section describes the application of the~~ 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 ~~(e.g. Seneviratne et al., 2010; Santanello et al., 2018)~~(Seneviratne et al., 2010; Santanello et a , 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 ~~upper panel of~~ Fig. ~~??~~2). Improving our understanding of soil moisture-precipitation coupling is important ~~, because this might~~ to improve precipitation predictions with numerical models. ~~As an illustrative example, we~~

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 ~~$\boldsymbol{C_i}[t] \in \mathbb{R}^{120\times180}$~~ $\boldsymbol{C_\ell}[t] \in \mathbb{R}^{120\times180}$, e.g. antecedent precipitation, that approximately fulfil the adjustment criteria from Sect. 2.1.2, at 120×180 input pixels (see Fig. 3)~~, and~~ . 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. ~~The~~ 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.

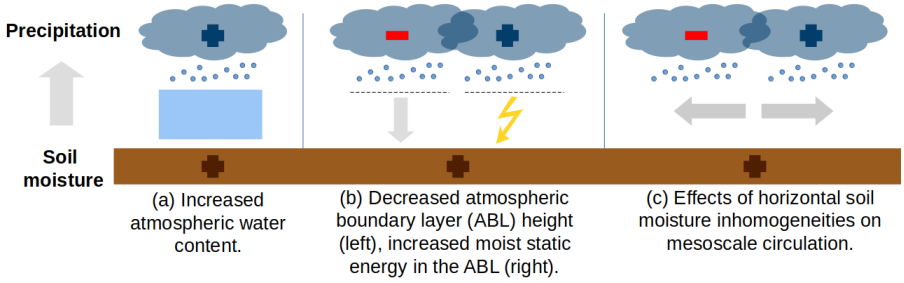

Figure 2. Concurring pathways of soil moisture-precipitation coupling. An increase in soil moisture can increase latent heat flux and decrease sensible heat flux at the land surface (Seneviratne et al., 2010). This can increase precipitation via an increase in atmospheric water content (a; Eltahir, 1998). 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).

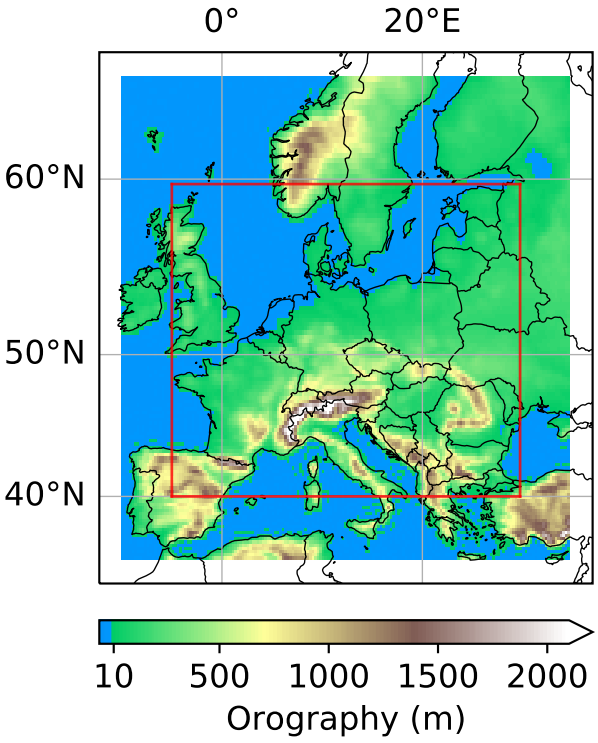

Figure 3. Input and target regions in the example of soil moisture-precipitation coupling. The colored region represents the 120×180 pixels input region, the red box the 80×140 pixels target region. Note that the offset between input and target region is 20 pixels on each side and distorted by the projection.

~~Section 3.1 provides details on the ERA5 data used for this example, Sect. 3.2 gives details on the considered loss function, DL model and general training implementation, and Sect. 3.3 details our choice of input variables. Finally, Sect. 3.4 describes the sensitivity analysis of the trained model.~~

3.1 Data

The data underlying our example are ERA5 hourly data (Hersbach et al., 2018) ~~, which are~~ constituting an atmospheric reanalysis of the past decades (1950 to today) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Reanalysis means ~~that they combine~~ 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 ~~lat-lon~~ latitude-longitude grid of 0.25 degrees (≈ 30 km). ~~Note that, in~~ In this study, soil moisture refers to the ERA5 variable "volumetric soil water in the upper soil layer (0-7 cm)~~. Note further that the~~". The target variable, precipitation ~~$\boldsymbol{P}[t + 4\ h]$~~ $\boldsymbol{P}[t + 4\ \mathrm{h}]$, represents an accumulation of precipitation over the ~~last hour.~~

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

~~From~~ As described in Sect. 2.2.1, ~~we have that~~ 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 ~~$\boldsymbol{C_i}[t]$~~$\boldsymbol{C_\ell}[t]$, i.e. by the function (cf. Eq. 6)

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

This holds true for the expected mean squared error ~~loss function,~~

$$L(\boldsymbol{x_1}, \ldots, \boldsymbol{x_N}) = \frac{1}{N}\sum_{i=1}^{N} \text{mean}((\boldsymbol{y_i} - \hat{\boldsymbol{y}}_{\boldsymbol{i}})^2))$$

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

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

~~that we use for this example.~~Here, ~~N is the number of training samples, $\boldsymbol{x_i} \in \mathbb{R}^{120\times180\times12}$ are the input samples, $\boldsymbol{y_i} \in \mathbb{R}^{80\times140}$ are the corresponding true precipitation fields and $\hat{\boldsymbol{y}}_{\boldsymbol{i}} \in \mathbb{R}^{80\times140}$ are the respective precipitation predictions of the model~~the mean operator denotes the mean over the 80×140 target pixels across Europe.

~~Further, the chosen~~ The employed DL model should be able to represent the presumably highly nonlinear function from Eq. 10. ~~As such a model, we choose a CNN~~ We choose a convolutional neural network (CNN; LeCun et al., 2015) whose architecture is inspired by the U-Net architecture ~~(Ronneberger et al., 2015) (see Fig. 4). When using this architecture, there are two concepts that one should be familiar with.~~(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 ~~certain neighborhood, the so called~~ surrounding region, the so-called receptive field. In ~~some cases, the neighborhood may comprise the entire input area. In these cases, we say that the receptive field is global. 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 ~~neighborhood.~~ surrounding region.

The second concept is that of translation invariance. ~~In the simplest case, translation~~ 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 ~~max pooling layers and transposed convolutions~~ the considered model architecture (Dumoulin

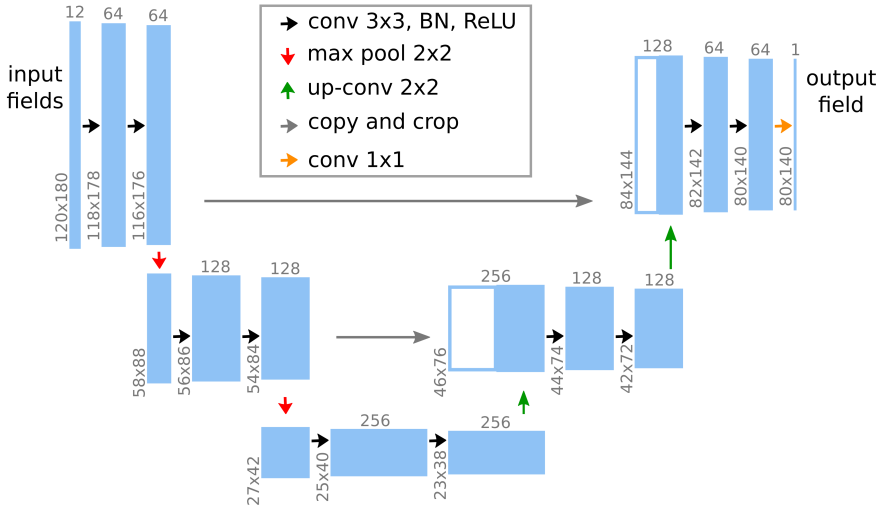

Figure 4. **Model architecture in the example of soil moisture-precipitation coupling**. The leftmost blue box represents the input to the model, which consists of 12 variables (including soil moisture) at the 120×180 input pixels (see Fig. 3). This input is passed through multiple sequential modules 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.

and Visin, 2016), the ~~model is actually~~ DL model is block translation invariant, i.e. the prediction at a target location (i, j) is ~~determined by the input variables in a 48×48 or 52×52 pixels receptive field and~~ not determined by a single function $\hat{f}$ for all target locations, but by one of 4×4 fixed functions $\hat{f}_{nk}, n, k = 1, \ldots, 4$ ~~. Here, the exact size and location of the receptive field and the choice of function $\hat{f}_{nk}, n, k \in \{1, \ldots, 4\}$, depends~~ depending on the values ~~$i \mod 4$ and $j \mod 4$.~~ $i \bmod 4$ and $j \bmod 4$.

Both concepts, receptive field and translation invariance, are ~~desirable~~ important features of CNNs~~as~~, because they counteract overfitting, i.e. making (nearly) perfect predictions on the training data but not generalizing to unseen data. However, ~~they also represent constraints to the model~~ both concepts constitute constraints that may prevent ~~it from being able to represent~~ CNNs from representing the function from Eq. 10. Indeed, ~~if the model is to learn this function, the~~ the translation invariance requires including ~~input variables~~ additional input variables $\{C_\ell\}_{\ell=1}^k$ that lead to spatial variability in soil moisture-precipitation coupling. We will ~~come back to this in the section on the choice of input variables.~~ discuss this in Sect. 3.3. Note that we can mostly ignore the general constraint of receptive fields, ~~i.e. that the prediction at a target location is fully determined by the input variables in a $\leq 52 \times 52$ pixels neighborhood,~~ because the lead time of the predictions is only 4 h and the receptive field is large enough ~~for the model~~ to take into account all relations between soil moisture and precipitation at that time scale.

Model architecture in the example of soil moisture-precipitation coupling. The model architecture is inspired by the U-Net architecture (Ronneberger et al., 2015). The input to the model are 12 variables (including soil moisture) at the 120×180 input pixels and the output is the precipitation prediction at the 80×140 target pixels (see Fig. 3).

Before starting to train Before training the model, we split our data in a into training, validation and test setsets. Due to the potential correlations between subsequent time steps, an entirely random split would lead to high correlations between samples in training, validation and test setsets. To achieve independence between samples belonging to different sets, we randomly chose 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 was held out is not used during the entire training and tuning process of the model.

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 (mse; see Eq. 11) 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

In this example, there are two aspects to consider when choosing input variables in addition to soil moisture. 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 our DL model to be able to approximate the function from 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. As the chosen DL model is translation invariant and we use rotation and flipping of the considered region as data augmentation during training, this requires the inclusion of input variables leading to spatial variability in soil moisture-precipitation coupling (see last section). Second, we want the) is a good approximation of the map

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

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 to form a sufficient set such that we can apply Theorem 1. Our choices of input variables are based on these two aspects and the descriptions of soil moisture-precipitation coupling in (Seneviratne et al., 2010; Santanello et al., 201 . Note however, that there is no *unique* translation of these studies into a particular choice of input variables . Section 4 discusses further analyses to assess to what extent our particular choices affect the results of the sensitivity analysisdescribed

~~in Sect. 3.4~~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
430 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.

~~To ensure that~~

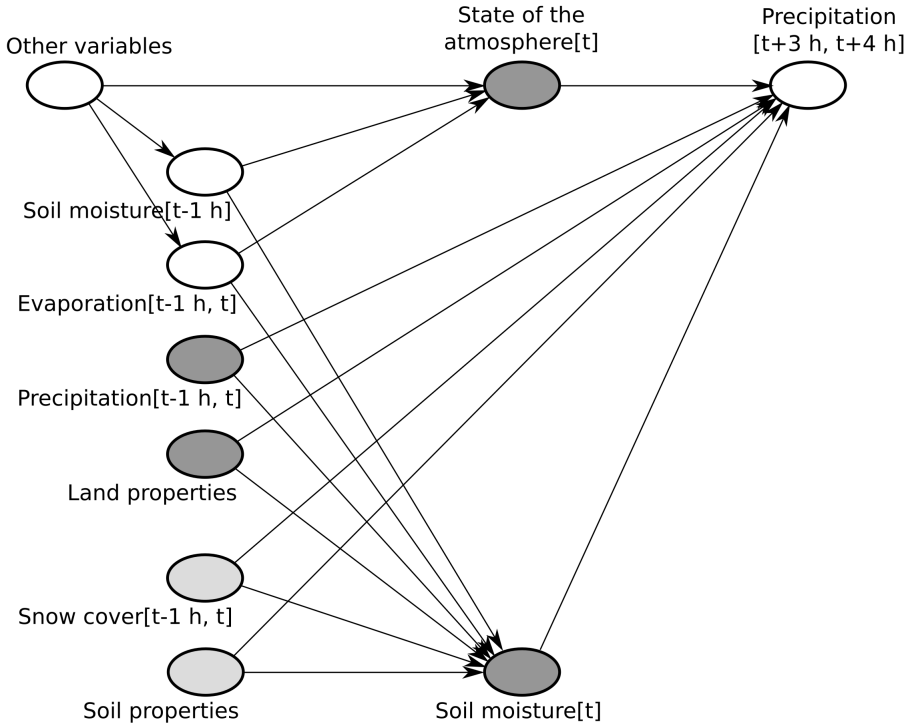

Figure 5. Causal graph in the example of soil moisture-precipitation coupling. The dark grey nodes represent the chosen input variables, while light grey nodes represent variables that are ignored in our analysis (see text). Land properties comprise the time-independent variables topography, land-sea 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 m height at time t, as well as wind at 100 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
435 to that from Eq. 12. Nevertheless, obtaining a good approximation of the map from Eq. 10 with our DL model ~~is able to learn the spatial variability of soil moisture-precipitation coupling, we have to either include latitude-longitude information or directly include the variables leading to spatial variability in soil moisture-precipitation coupling.We decided for the latter~~

because this allows the model to easily generalize between different locations in would be difficult due to the considered region (and in principle also outside that region). If instead we included strong correlation between $\boldsymbol{SM}[t-1\text{ h}]$ and $\boldsymbol{SM}[t]$. Furthermore, the strong correlation between evaporation$[t-1\text{ h}, t]$ and evaporation$[t, t+3\text{ h}]$ may prevent us from identifying any causal effect of $\boldsymbol{SM}[t]$ on $\boldsymbol{P}[t+4\text{ h}]$, because evaporation$[t, t+3\text{ h}]$ is a direct descendant of $\boldsymbol{SM}[t]$ on every causal path from $\boldsymbol{SM}[t]$ to $\boldsymbol{P}[t+4\text{ h}]$ (cf. motivation of the second adjustment criterion in Sect. 2.1.2). Therefore, we decided to exclude $\boldsymbol{SM}[t-1\text{ h}]$ and evaporation$[t-1\text{ h}, t]$. Nevertheless, this leads to unblocked non-causal paths between $\boldsymbol{SM}[t]$ and $\boldsymbol{P}[t+4\text{ h}]$ via these variables (e.g. $\boldsymbol{SM}[t] \leftarrow \boldsymbol{SM}[t-1\text{ h}] \rightarrow$ state of the atmosphere$[t] \rightarrow \boldsymbol{P}[t+4\text{ h}]$). 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 $\boldsymbol{SM}[t-1\text{ h}]$ and evaporation$[t, t-1\text{ h}]$ 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 ~~soil moisture-precipitation coupling function~~ mapping for each location ~~. Further,~~ (i, j), and data augmentation in form of flipping and rotation of the region would not ~~make sense. Specifically, we included land-sea mask, fraction of high and low vegetation cover,~~ be useful. Instead, we include short- and long-wave radiation at the land surface$[t]$~~, 2 temperature$[t]$ and 2 dew point temperature$[t]$ to take into account spatial differences in the evaporation process. Note, that the addition $[t]$ means, that the variable is considered at the same time as the soil moisture input variable, while for example land-sea mask and vegetation cover in the considered ERA5 data are constant in time. Further, U and V components of the wind$[t]$ are included as they determine spatial differences in the distribution of locations influenced by soil moisture changes.For example, if at some location, the wind blows mainly westward, mainly precipitation at westward locations will be affected by soil moisture changes. Finally, topography is included, because it dominates spatial differences in precipitation.~~ . 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 ~~second aspect to consider when choosing input variables in addition to soil moisture is that we want them to form a sufficient set such that we can apply Theorem 1. The first condition for a sufficient set is that we do not include any input variable that is a descendant of soil moisture$[t]$, i.e. that is in some way causally influenced by soil moisture$[t]$. We achieve this by not including any input variable for a time $\hat{t} > t$, nor variables that might instantaneously be affected by soil moisture$[t]$, e.g. evaporation$[t]$.~~

~~The second condition for a sufficient set is that the additional input variables block all paths between soil moisture$[t]$ and precipitation$[t+4\text{ h}]$ that contain an edge pointing to soil moisture$[t]$. One option to achieve this is to include all parents of soil moisture$[t]$, i.e. all variableswhich have a direct impact on soil moisture$[t]$, that also affect precipitation$[t+4\text{ h}]$. Most~~

~~important of these variables may be antecedent precipitation and antecedent soil moisture. Antecedent precipitation increases soil moisture$[t]$ and at the same time is correlated with precipitation$[t + 4\text{ h}]$ (e.g. when precipitation occurs in large-scale synoptic weather systems). This leads to a non-causal correlation between soil moisture$[t]$ and precipitation$[t + 4\text{ h}]$. By including antecedent precipitation as input variable, or, in other words, conditioning on antecedent precipitation, we can exclude this correlation from our analysis. On the other hand, antecedent soil moisture$[t - \Delta]$ (for some small Δ) affects soil moisture$[t]$ and at the same time precipitation$[t + 4\text{ h}]$, thereby leading to a non-causal correlation between soil moisture$[t]$ and precipitation$[t + 4\text{ h}]$. Conceptually, if we did not take into account antecedent soil moisture, this would disturb the time scale of our analysis, because from a change in soil moisture$[t]$, the model would expect a change in antecedent soil moisture$[t - \Delta]$, which would also affect expected precipitation$[t + 4\text{ h}]$. However, rather than directly including antecedent soil moisture as input variable, we decided to include variables that block the paths from antecedent soil moisture to precipitation. The motivation for this is twofold. First, we already included many of these variablesanyway,~~ 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. ~~2 temperature$[t]$ and 2 dew point temperature$[t]$, to take into account the spatial variability of~~ 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. ~~Second, while including antecedent soil moisture$[t - \Delta]$ is valid in theory, in practice, correctly learning the map~~

$$(\boldsymbol{SM}[t], \boldsymbol{SM}[t-\Delta], \{\boldsymbol{C_i}[t]\}_{i=1}^{k-1}) \rightarrow \mathbb{E}[\boldsymbol{P}[t+4\text{ h}] | \boldsymbol{SM}[t], \boldsymbol{SM}[t-\Delta], \{\boldsymbol{C_i}[t]\}_{i=1}^{k-1}]$$

~~from Eq. 10 may be difficult for the DL model due to the strong correlation between $\boldsymbol{SM}[t]$ and $\boldsymbol{SM}[t - \Delta]$.~~

~~Our final choice of input variables is summarized by the~~ 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 in Fig. 5. The dark grey nodes represent the input variables chosen in addition to soil moisture$[t]$ and pink paths represent the causal paths that are blocked by this choice. Note that there are many more variables related to soil moisture-precipitation coupling and further causal paths, and our choice of input variables only *approximates* a sufficient set. Section 4 discusses further analyses to assess to what extent this affects the results of the sensitivity analysis described in Sect. 3.4.~~ graph clearly conveys to other scientists the assumptions underlying a specific application of the proposed methodology.

~~**Causal graph for studying soil moisture-precipitation coupling.** The dark grey nodes represent the input variables chosen in addition to soil moisture$[t]$ and pink paths represent the causal paths that are blocked by this choice. The effects of neglecting other potentially important variables and paths are discussed in Sect. 4.~~

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 (~~see~~ cf. Sect. 2.2.2). ~~As local effect or local soil moisture-precipitation coupling, we define the impact~~ We define the causal effect of a soil moisture change at a pixel ~~$(i,j) \in \{1,\dots,80\} \times \{1,\dots,140\}$ in the 80 × 140 pixels target region~~ (i,j) on precipitation at the very same pixel ~~(i,j)~~as local effect or local soil moisture-precipitation coupling. Accordingly, we consider for each pixel (i,j) in the target region the partial derivative

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

where $\boldsymbol{p}_{ij}$ denotes the precipitation prediction of the DL model for pixel (i,j), and $\boldsymbol{SM}$ and ~~$\{\boldsymbol{C_n}\}_{n=1}^k$~~ $\{\boldsymbol{C_\ell}\}_{\ell=1}^k$ are the input variables to the model. ~~To obtain the average impact of a change in local soil moisture on local precipitation, we average~~ We average these derivatives over all input samples ~~$(\boldsymbol{SM}, \{\boldsymbol{C_n}\}_{n=1}^k)$~~ $(\boldsymbol{SM}, \{\boldsymbol{C_\ell}\}_{\ell=1}^k)$ from the test set ~~, which we denote by $\overline{s^{loc}}_{ij}$~~denoted by $\overline{q^{loc}}_{ij}$.

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

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

~~Again, these derivatives are averaged over all input samples from the test set to obtain the average impact of a change in local soil moisture on regional precipitation, which we denote by $\overline{s^{reg}}_{ij}$.~~ Note that most of the derivatives in the sum are zero, ~~as for instance~~ because e.g. a change in soil moisture ~~[t]~~ in Great Britain at time t does not affect precipitation ~~[t + 4 h] in Italy~~ in Italy four hours later. Outside of a 52×52 pixels ~~neighborhood~~surrounding region, this is enforced by the architecture of the DL model (~~see the concept of receptive fields explained in~~ cf. Sect. 3.2)~~and within the 52 × 52 pixels neighborhood, this~~ , 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^{reg}_{ij} over all input samples from the test set.

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

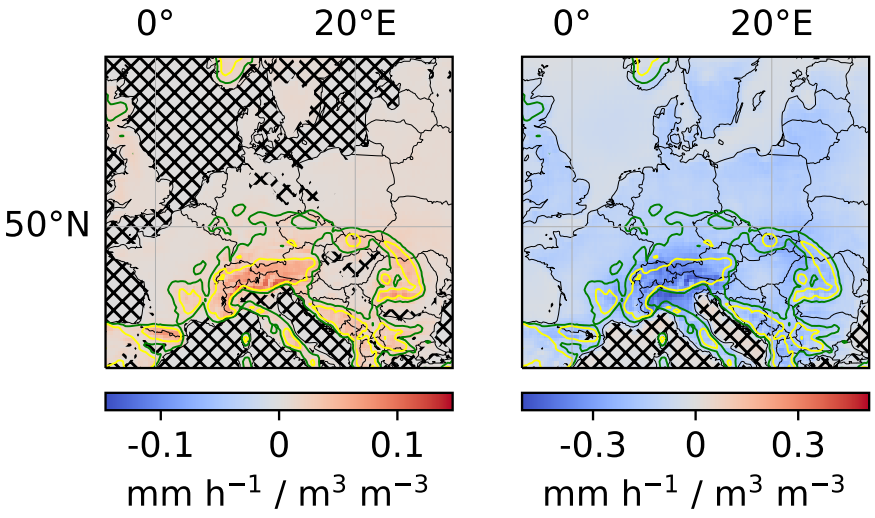

Figure 6. ~~**Local and regional soil moisture-precipitation coupling**~~**Local and regional soil moisture-precipitation couplings**. Left: Impact of local soil moisture changes (m^3 water $\cdot\, \mathrm{m}^{-3}$ soil) on local precipitation ($\mathrm{mm\,h}^{-1}$) for each pixel in the target region (in the text denoted by $\overline{s^{loc}}_{ij}\overline{q^{loc}}_{ij}$). Right: Impact of local soil moisture changes on regional precipitation for each pixel in the target region (in the text denoted by $\overline{s^{reg}}_{ij}\overline{q^{reg}}_{ij}$). ~~Note that the unit $\mathrm{mm\,h}^{-1}$ for precipitation always refers to some area.~~ For better comparability of local and regional values, ~~it~~ 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 ~~randomness~~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.

4 ~~Further analyses to assess the correctness of obtained results~~

3.1 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., 201
. 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.

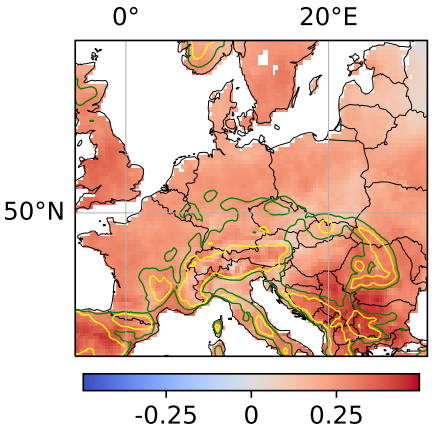

Figure 7. Linear correlation coefficient of local soil moisture and regional precipitation. For each location, the linear correlation coefficient of soil moisture $\boldsymbol{SM}[t]$ at the location and subsequent precipitation $\boldsymbol{P}[t+4\,\mathrm{h}]$ summed over the 15×15 pixels surrounding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

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., 201 . This approach allows to evaluate the effects of soil moisture changes on precipitation 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.

4 Additional analyses to verify the results

~~There are several things that may go wrong in the proposed methodology and lead to results that do not reflect the causal impact of~~ 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 $\boldsymbol{Y}$ ~~but spurious correlations. For example, our input variables might not approximate a sufficient set "well enough" due to an incomplete or incorrect underlying causal graph;~~

~~our DL model or training procedure might not be suitable to learn the function from Eq. 6; or $\boldsymbol{X}$ might simply not affect $\boldsymbol{Y}$. In this section, we propose several further 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 (Sect. 4.1); whether they reflect more than specific (known) correlations (Sect. 4.2); and whether they actually reflect causal rather than (potentially unknown) spurious correlations (Sect. 4.3). Finally, we propose some further sanity checks in Sect. 4.4. We illustrate these analyses with the example of soil moisture-precipitation coupling.~~ 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.

~~Note that, on its own, the performance of the model on the test set is no good indicator for the correctness of obtained results. On the one hand, the performance might be "good", although the learned $\boldsymbol{X}$-$\boldsymbol{Y}$ coupling is wrong, e.g. when the good performance is due to mere correlations between $\boldsymbol{X}$ and $\boldsymbol{Y}$, or due to the other input variables $\boldsymbol{C_i}$.On the other hand, the performance might be "bad", although the learned $\boldsymbol{X}$-$\boldsymbol{Y}$ coupling is correct. Consider for example~~ 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 ~~$\boldsymbol{y} = \boldsymbol{x} + 1000\boldsymbol{c}$, where the values of~~ $\boldsymbol{Y} = \boldsymbol{X} + 1000 \cdot \boldsymbol{C}$ (where $\boldsymbol{X}$ and $\boldsymbol{C}$ vary in similar ranges~~. In this case, to determine the correct $\boldsymbol{X}$-$\boldsymbol{Y}$ coupling,)~~, the adjustment criteria from Sect. 2.1.2 imply that it suffices to ~~train a DL model to predict $\boldsymbol{Y}$ given~~ consider $\boldsymbol{X}$ ~~. However, the performance of this DL model cannot be good because~~ 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 $\boldsymbol{C}$.

~~Note however that the performance on the test set *can* be an indicator for the correctness of obtained results, when it is for example compared to the performance of the model for permuted values of~~ 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 $\boldsymbol{X}$ ~~. We detail this in~~ 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. ~~4.2. Solely for reference , we note that the~~ 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 ~~(mse) with respect to the normalized target variables (~~on the test set (target variable normalized to mean of 0 and standard deviation of 1 on the training set) ~~on the test set,~~

~~averaged over the ten considered instances of the DL model, is 0.60, whereas it is 1.54 for~~ for our application to soil moisture-precipitation coupling: it is 0.60 for the DL model. For a persistence prediction~~(~~, i.e. ~~for~~ when predicting the input ~~field~~ $\boldsymbol{P}[t]$ as target ~~field~~ $\boldsymbol{P}[t+4\text{ h}]$~~)~~, which is a simple baseline prediction, it is 1.54.

4.1 ~~Are the obtained results statistically significant? I.e. do they reflect more than random correlations or artifacts of the DL training procedure?~~Statistical significance

~~Given some~~ To test whether results obtained with the proposed methodology are statistically significant, i.e. represent more than random correlations between $\boldsymbol{X}$ ~~and~~ and $\boldsymbol{Y}$ ~~sensitivity $s \in \mathbb{R}$ obtained by the sensitivity analysis described in Sect. 2.2.2 (e.g. $s_{i_1 i_2}$ from Eq. ?? for some tuple (i_1, i_2)), one might wonder, whether its value really reflects that $\boldsymbol{X}_{i_2}$ contains information on $\boldsymbol{Y}_{i_1}$, or whether it is random or an artifact of the DL training procedure. To test this ,~~ 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 ~~s as a random variable $s : \Omega \to \mathbb{R}$~~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

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

~~Now, the null hypothesis is that the value of s does not reflect any information in~~ 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 $\boldsymbol{X}$ ~~on~~ and $\boldsymbol{Y}$ ~~, but is random or an artifact of the training procedure~~ in the training data, or random artifacts of the procedure for training the DL model. To test this hypothesis, we create ~~a sample ω_0^1~~ m samples $\omega_0^1, \ldots, \omega_0^m$ of Ω ~~under the null hypothesis. To that purpose, we randomly permute~~ by the above described procedure of permuting $\boldsymbol{X}$ ~~(but not $\boldsymbol{Y}$, nor the additional input variables $\boldsymbol{C_t}$!) in the training set, in a way that breaks all correlations between $\boldsymbol{X}$ and $\boldsymbol{Y}$ (e.g. in the example of soil moisture-precipitation coupling, we permute soil moisture temporally and spatially), while preserving relations between $\boldsymbol{C_t}$~~ and ~~$\boldsymbol{Y}$, and preserving the general distribution of $\boldsymbol{X}$. Further, we randomly initialize a new instance of the~~ randomly initializing the weights of separate instances of the considered DL model. ~~Next, we train this new instance of the DL model on the modified training set and obtain a sensitivity $s_0^1 \in \mathbb{R}$. We repeat this process k times to obtain k samples$\omega_0^1, \ldots, \omega_0^k$ and corresponding sensitivities $s_0^1, \ldots, s_0^k$. Given large k, we could~~ 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 ~~sensitivity s lay~~ value r lies outside the middle $100\,\% - \alpha$ of the values ~~$s_0^1, \ldots, s_0^k r_0^1, \ldots, r_0^m$~~, i.e. if

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

However, because we have to train ~~k models~~ m DL models for this analysis, it may not be feasible to choose ~~k~~ m large enough to get reasonable approximations of ~~the~~ these percentiles. In this case, we propose ~~to compute~~ computing the mean μ and standard deviation σ of the values ~~$s_0^1, \ldots, s_0^k$~~, $r_0^1, \ldots, r_0^m$, ~~*assume*~~*assuming* a normal distribution of ~~s_0, and reject~~ $\hat{r}$ under the null hypothesis, and rejecting the null hypothesis at significance level α if ~~s~~ r is not in the middle $100\,\% - \alpha$ of the distribution $N(\mu, \sigma)$, i.e. if

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

~~While this test is enough to show that the value of s reflects some information in $\boldsymbol{X}$ on $\boldsymbol{Y}$, and is not solely random or an artifact of the training procedure, Sect. 4.2 details how it might be taken a step further. Namely, in Sect. 4.2, we consider the null hypothesis that "the value of s is random or an artifact of the training procedure *or* reflects solely specific correlations $c_1, \ldots, c_c$ between $\boldsymbol{X}$ and $\boldsymbol{Y}$". Rejection of this hypothesis implies rejection of the hypothesis that "the value of s is random or an artifact of the training procedure". Therefore, we limit Fig. 6 to showing the results from Sect. 4.2, while the results from this section are not shown.~~

4.2 ~~Do the obtained results reflect more than specific correlations? E.g. more than seasonality and the correlation between soil moisture and topography?~~

4.2 Known spurious correlations

~~Given some~~ As mentioned above, the proposed methodology identifies the exact causal effect of $\boldsymbol{X}$ ~~on~~ $\boldsymbol{Y}$ ~~sensitivity $s \in \mathbb{R}$ obtained by the sensitivity analysis described in Sect. 2.2.2 (e.g. $\overline{s_{i_1 i_2}}$ from Eq. **??** for some tuple (i_1, i_2)), one might wonder, whether its value reflects more than some specific correlations $c_1, \ldots, c_c$~~ 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, ~~when considering our results for soil moisture-precipitation coupling in Fig. 6, one might wonder whether they reflect solely potential correlations due to seasonality~~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~~), and the combination of soil moisture-topography correlation and topography-precipitation correlation. To test this, we propose two permutation-based approaches, which apply whenever~~. 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.

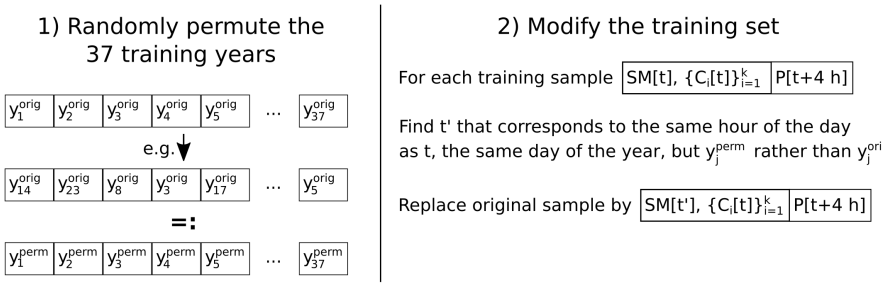

Figure 8. Modification of the training data for the year-wise permutation of $SM[t]$. The modification of the test data works analogously.

The first proposed analysis is identical to the analysis described in Sect. 4.1 except that $\boldsymbol{X}$ ~~can be permuted~~ in the training data is not permuted randomly, but such that the ~~correlations $c_1, \ldots, c_c$ are preservedwhile other correlationsbreak. For example , when considering~~ 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 ~~randomly permute the soil moisture years, thereby preserving the seasonality correlation and the correlation~~ 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 ~~(and of course between topography and precipitation)~~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.

~~In the first approach, we consider k (in our case $k = 10$)~~ 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 ~~DL model which were trained with different random weight initializations. We compute the mean squared error (mse)of these instances~~ 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~~and obtain k values $\mathrm{mse}_1, \ldots, \mathrm{mse}_k$~~, obtaining m values $l_1, \ldots, l_m \in \mathbb{R}$. Next, ~~we permute the soil moisture input years~~ for each model instance separately, we randomly permute $\boldsymbol{X}$ in the test ~~set (as there are only two test years, this corresponds to switching both years)~~ data such that the considered spurious correlations are preserved, and compute the ~~k corresponding values $\overline{\mathrm{mse}}_1, \ldots, \overline{\mathrm{mse}}_k$~~value of the loss function on the modified test set, obtaining m values $l_1^{\mathrm{perm}}, \ldots, l_m^{\mathrm{perm}} \in \mathbb{R}$. Finally, we use a permutation test (Hesterberg, 2014) to test if the expected value of the loss function is smaller on the ~~null hypothesis that the expected mse is worse or equal when considering the~~

original test set than ~~when considering the test set with permuted soil moisture years.~~ 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 ~~, the null hypothesis was rejected~~ 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 % ~~, indicating~~ that the model learned ~~more than seasonality and soil moisture-topography correlation. Note that~~ 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 ~~test~~analysis, it may be ~~harmful~~ limiting that there are only two test years in ~~our case~~ this example and thus only one possible permutation of years apart from the original one. Therefore, we repeated the ~~test and permuted the soil moisture input time steps~~ analysis and permuted soil moisture in the test ~~set several times completely randomly~~ data completely randomly in time. While this ~~breaks potential correlations due to~~ does not preserve correlations between soil moisture and seasonality, it still preserves the correlation between soil moisture and topography. ~~Again,~~ Furthermore, it ensures the validity of the ~~null hypotheses of worse or equal mse when considering the original test set was rejected~~ analysis as there are a lot of possible permutations. In this case, the analysis indicates at a confidence level of 99 % ~~, indicating~~ that the model learned ~~more than the correlation between soil moisture and topography.~~

~~The second approach that we propose to answer the question if the obtained results reflect more than specific correlations $c_1, \ldots, c_c$ requires to train k (in our case $k = 10$) new models. In addition to this question, it may also answer the question from Sect. 4.1, namely, if the obtained results reflect more than random correlations or artifacts of the training procedure, and are statistically significant. In particular, given some $\boldsymbol{X}$-$\boldsymbol{Y}$ sensitivity $s \in \mathbb{R}$ obtained by the sensitivity analysis described in Sect. 2.2.2 (e.g. $\overline{s_{i_1 i_2}}$ from Eq. **??** for some tuple (i_1, i_2)), we consider the null hypothesis "the value of s is random or an artifact of the training procedure *or* reflects solely specific correlations $c_1, \ldots, c_c$ between $\boldsymbol{X}$ and $\boldsymbol{Y}$". In the example of~~ useful information in terms of predictive performance on soil moisture-precipitation coupling, ~~we consider again the potential correlation due to seasonality, and the combination of soil moisture-topography correlation and topography-precipitation correlation. The sensitivity s represents either local or regional soil moisture-precipitation coupling for some location (i, j) in the considered region (i.e. either $\overline{\boldsymbol{s}^{\boldsymbol{loc}}}_{ij}$ or $\overline{\boldsymbol{s}^{\boldsymbol{reg}}}_{ij}$ from Sect. 3.4). To test the null hypothesis, we proceed similarly to Sect. 4.1. Namely, we create a sample ω_0^1 of Ω under the null hypothesis by randomly permuting the 37 training years for the soil moisture input variable, thereby preserving the~~ apart from the correlation between soil moisture and topography~~(and of course between topography and precipitation) and potential correlations due to seasonality. Further, we randomly initialize a new instance of the DL model. Next, we train this new instance of the DL model on the modified training set and obtain a sensitivity $s_0^1 \in \mathbb{R}$. We repeat this process k times to obtain k samples $\omega_0^1, \ldots, \omega_0^k$ and corresponding sensitivities $s_0^1, \ldots, s_0^k$, compute the mean μ and standard deviation σ of these k values and test if~

$s \in [\text{percentile}(N(\mu, \sigma), \alpha/2), \text{percentile}(N(\mu, \sigma), 100\,\% - \alpha/2)]$,

where we set $\alpha = 5\,\%$. Note that it suffices to train k models to test the null hypothesis for local and regional coupling, respectively, and for all locations (i, j). This is because the k models trained on the modified training sets $\omega_0^1, \ldots, \omega_0^k$ do not only yield k estimates of $\overline{s^{loc/reg}}_{ij}$ for a single location (i, j), but yield k estimates of $\overline{s^{loc/reg}}_{ij}$ for each location (i, j).

The results of this analysis are illustrated in Fig. 6. In particular, missing hatching indicates that the null hypothesis at a location was rejected, i.e. that the value of local/ regional coupling at this location Note that even if the first analysis indicates that some result reflects more than randomness, artifacts of the training procedure, seasonality, and the correlation between soil moisture and topography. On the other hand, hatching indicates that the null hypothesis could not be rejected. While the hatching indicates that our results are significant and do not only reflect soil moisture-topography correlation or seasonality, we are not sure why most of the ocean in the regional coupling is not hatched although soil moisture at these locations is set to one for all time steps. We suspect that it is related to the fact that we set soil moisture to constant one for all ocean locations while soil moisture is smaller than 0.75 for all non-ocean locations. Missing variation of soil moisture values around the value 1 chosen for ocean locations could lead to the DL model simply not caring about soil moisture "sensitivities" for ocean locations. We welcome any discussion on this.

Note that, from these analyses, we cannot conclude that the obtained results are not partly due to the correlations $c_1, \ldots, c_C$, but only that they are not *entirely* due to these correlations, randomness, or artifacts of the training procedurethe 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 Do the obtained results reflect (potentially unknown) Further spurious correlations?

In the last two sections, we already proposed approaches to identify results which only reflect random correlationsor artifacts of DL model training, or specific (known) correlations . To assess, whether they reflect potentially unknown spurious correlations rather than a causal impact of X on Y, e.g. due to an incomplete or incorrect underlying causal graphprevious 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). The concept is to train 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 relations that results obtained from original and variant models learn 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 ~~in general~~ do not vary in time or space. Then, if the couplings in Fig. 6 reflect the causal ~~impact~~ 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 ~~and~~ or second half of the training years, ~~respectively,~~
- data from June, July ~~and August, respectively,~~ or August, or
- the left ~~and right half , respectively,~~ 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 ~~Figures~~Figs. A1 to A3 show the local and regional couplings obtained from the different model instances trained on the listed training subsets. As ~~it should be if~~ expected for the case that all instances learned the causal ~~impact~~ effect of soil moisture on precipitation, all couplings are very similar to the ones shown in Fig. 6. Note however that ~~the variant approach only tests a necessary condition, not a sufficient condition. For example, if Fig. 6 reflected spurious correlations and these spurious correlations did not vary in the different training subsets listed above, we would not be able to identify them with this approach.~~ this does not guarantee that they show causal relations.

4.4 ~~Further~~ Task-specific sanity checks

To further assess the correctness and increase trust in results obtained from the proposed methodology, ~~one might~~ we propose to perform further, task-specific sanity checks. ~~In the~~ For instance, in our example of soil moisture-precipitation coupling, ~~for instance,~~ precipitation $\boldsymbol{P}$ can be partitioned into convective precipitation $\boldsymbol{P_{con}}$ (occurring at spatial scales smaller than the ~~grid box~~spatial resolution of the numerical model) and large-scale precipitation $\boldsymbol{P_{ls}}$ (occurring at larger spatial scales), such that $\boldsymbol{P = P_{con} + P_{ls}}$. 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 ~~simply~~ replacing $\boldsymbol{P}$ by $\boldsymbol{P_{con}}$ and $\boldsymbol{P_{ls}}$, respectively, and compared the sum of the obtained couplings with Fig. 6. Appendix ~~Figure~~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 ~~$\boldsymbol{SM}$ input variable~~ input variable $\boldsymbol{SM}$ by $\boldsymbol{E}$. Appendix ~~Figure~~Fig. A7 shows the product of local $\boldsymbol{SM}$-$\boldsymbol{E}$ and local and regional $\boldsymbol{E}$-$\boldsymbol{P}$ ~~coupling~~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.

5 ~~Comparison to linear correlation~~

~~For a deeper analysis of~~

4.1 Control experiment

As a simple control experiment for the proposed methodology ~~, it would be interesting to perform an ablation study, i.e. repeat the above experiments with a different loss function, a less powerful statistical model, without input variables leading to spatial variability in soil moisture-precipitation coupling, and without input variables that approximate a sufficient set, respectively. Here, we limit ourselves to a comparison with the results obtained from a simple linear correlation analysis. In particular, for each location in the considered target region, Fig. 7 shows the linear correlation between soil moisture $\boldsymbol{SM}[t]$ at the location and subsequent precipitation~~ and analyses, we replaced the target variable $\boldsymbol{P}[t + 4\,\mathrm{h}]$ ~~summed over the 15×15 pixels neighborhood of the location. Being conceptually similar to our analysis of regional soil moisture-precipitation coupling in the right panel of Fig. 6, the linear correlation analysis assumes linearity of relations between local soil moisture and regional precipitation, and completely neglects the discrepancy between causality and correlation. The obtained regional~~ 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" in Fig. 7 then also differs entirely from the coupling in the right panel of Fig. 6, stressing the importance of considerations made in the proposed methodology~~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.

~~**Linear correlation between local soil moisture and regional precipitation.** For each location, it is shown the linear correlation between soil moisture $\boldsymbol{SM}[t]$ at the location and subsequent precipitation $\boldsymbol{P}[t + 4\,\mathrm{h}]$ summed over the 15×15 pixels neighborhood of the location. Compare to right panel of Fig. 6.~~

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 ~~proposed~~ methodology is based on the recent idea of training and analyzing a DL model to gain new scientific insights ~~on~~ into the relations between input and target variables. It extends this idea by combining it with ~~insights~~ concepts from causality research. ~~Summarizing~~ A crucial aspect in the proposed methodology ~~, given a complex relation between two variables , for example soil moisture-precipitation coupling, we train a DL modelto predict one variable given the other, and perform a sensitivity analysis of the trained model. To achieve that the DL model actually learns the causal impact of~~ is the choice of additional input variables for the DL model. This choice requires prior knowledge on which variables are relevant to the ~~respective input variable on the target variable, the loss function, DL model and additional input variables are chosen carefully~~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~~itself, we proposed several further~~ , 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 ~~some further~~ sanity checks for the obtained results. While ~~these~~ the analyses cannot guarantee the correctness of obtained results, ~~and developing further analyses is desirable,~~ we believe that the proposed analyses provide a solid indication of the correctness of obtained results. ~~Note that studies based on numerical simulations, which rely on many assumptions in the numerical model, and other statistical approachescannot guarantee correctness either~~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 ~~further~~ 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 ~~a~~ particularly strong local and regional ~~coupling~~ 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 ~~general~~ methodology. An ~~extensive~~ extension and discussion of our results on soil moisture-precipitation coupling in terms of physical processes ~~and related studies will follow in a second paper~~are subject of a future study.

~~We believe that, harnessing the great power and flexibility of DL models, the proposed methodology may yield new scientific insights into complex, e.g. nonlinear and non-local, mechanisms in the Earth system.~~

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.

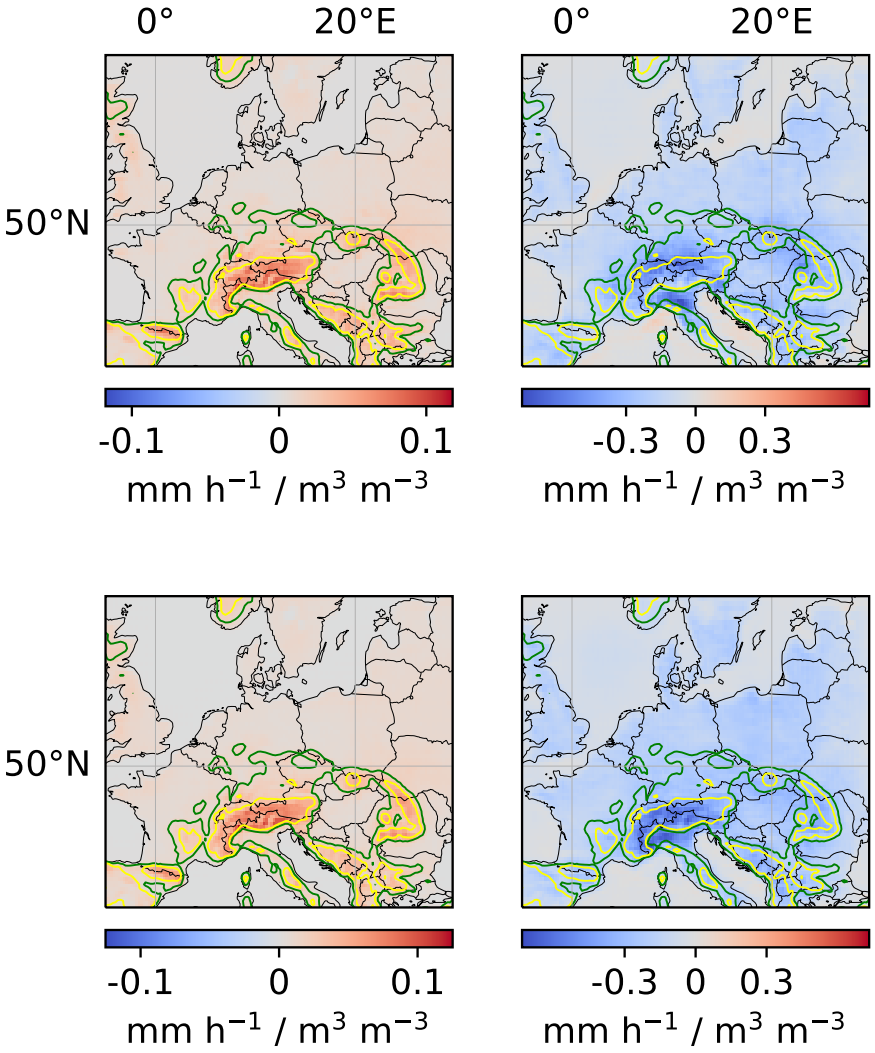

Figure A1. ~~Local and regional soil moisture-precipitation coupling for models trained on the first and second half of the training years, respectively~~Local and regional soil moisture-precipitation couplings for models trained on the first and second half of the training years, respectively. Left column: local ~~coupling~~couplings. Right column: regional ~~coupling~~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).

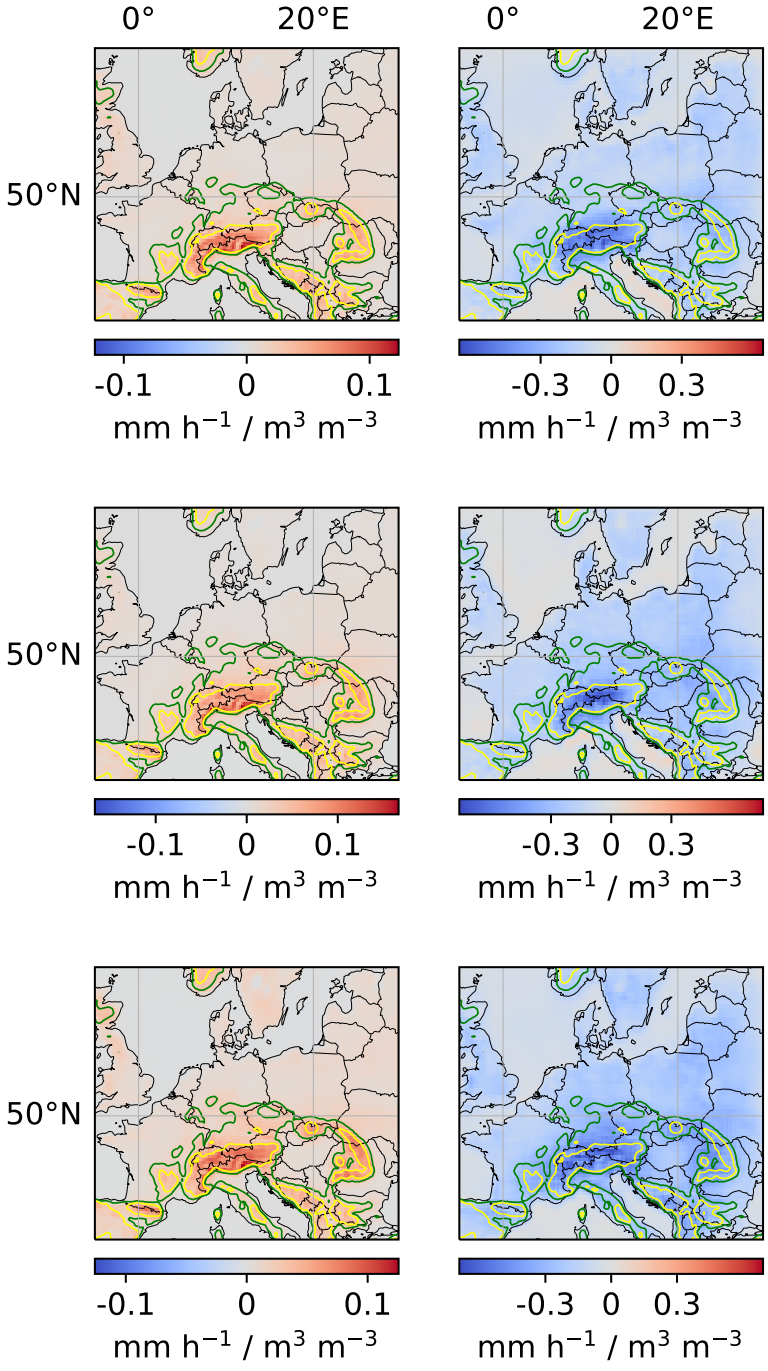

Figure A2. ~~Local and regional soil moisture-precipitation coupling for models trained only on data from June, July and August, respectively~~**Local and regional soil moisture-precipitation couplings for models trained only on data from June, July and August, respectively**. Left column: local ~~coupling~~couplings. Right column: regional ~~coupling~~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.

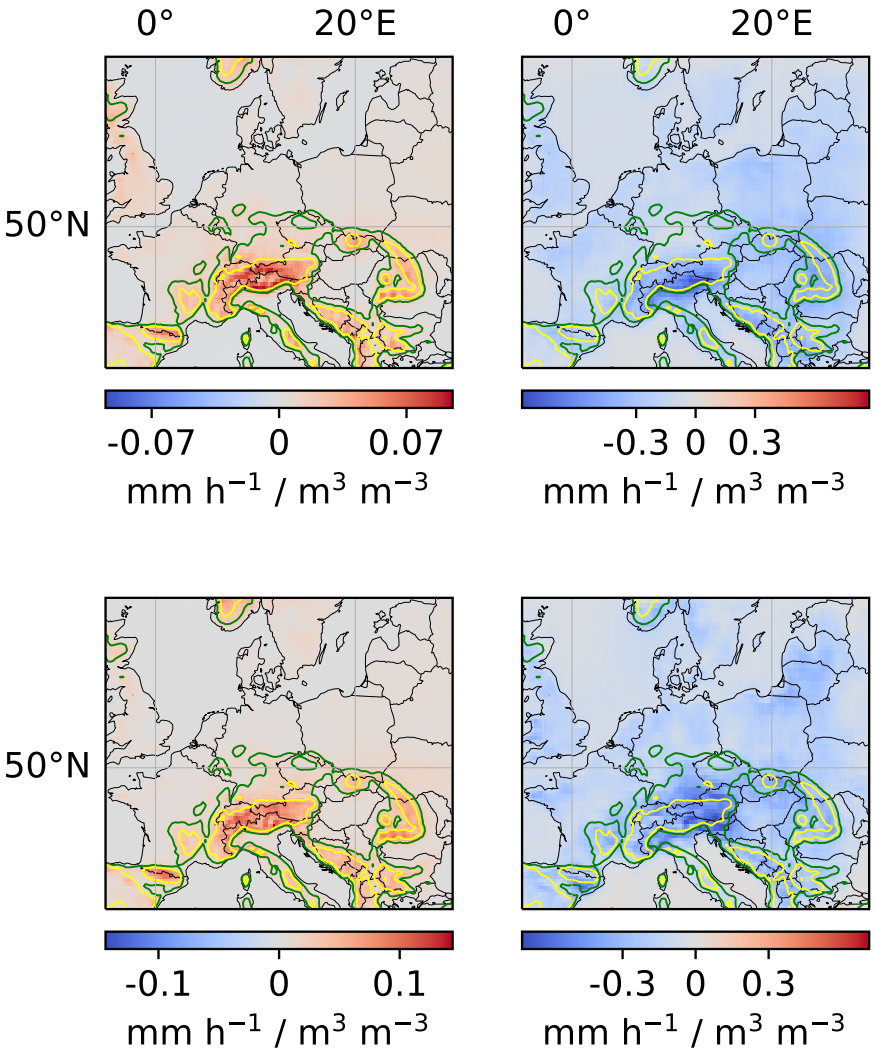

Figure A3. ~~Local and regional soil moisture-precipitation coupling for models trained on the left and right half of the considered region, respectively~~**Local and regional soil moisture-precipitation couplings for models trained on the left and right half of the considered region, respectively**. Left column: local ~~coupling~~couplings. Right column: regional ~~coupling~~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 ~~, while~~ the models were trained only on the left and right half, respectively, but the ~~CNN~~ model architecture allows to compute local and regional couplings for the entire region~~, which is shown here~~.

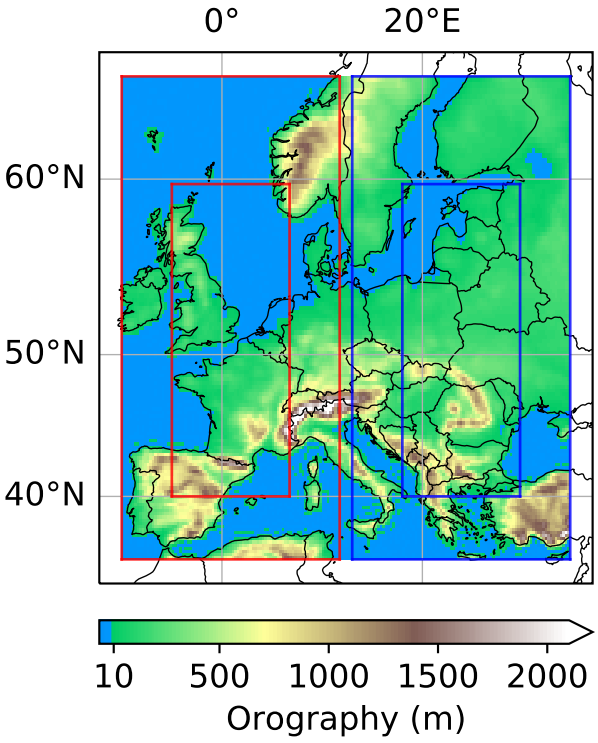

Figure A4. Location variant tasks. The input region was divided in a left and a right input region with corresponding target regions (indicated by the red and blue boxes).

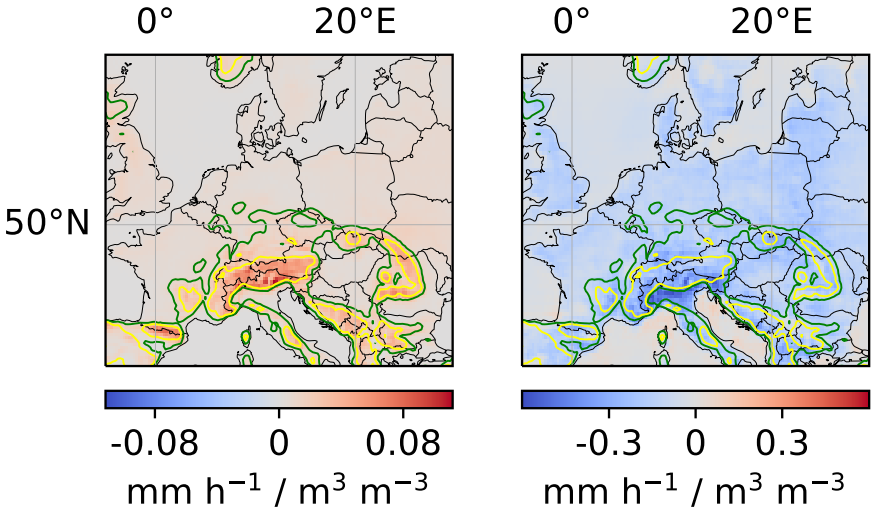

Figure A5. Sum of local and regional soil moisture-convective precipitation and soil moisture-large-scale precipitation couplings. Left: sum of local couplings. Right: sum of regional couplings. See Appendix Fig. A6 for soil moisture-convective precipitation and soil moisture-large-scale precipitation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

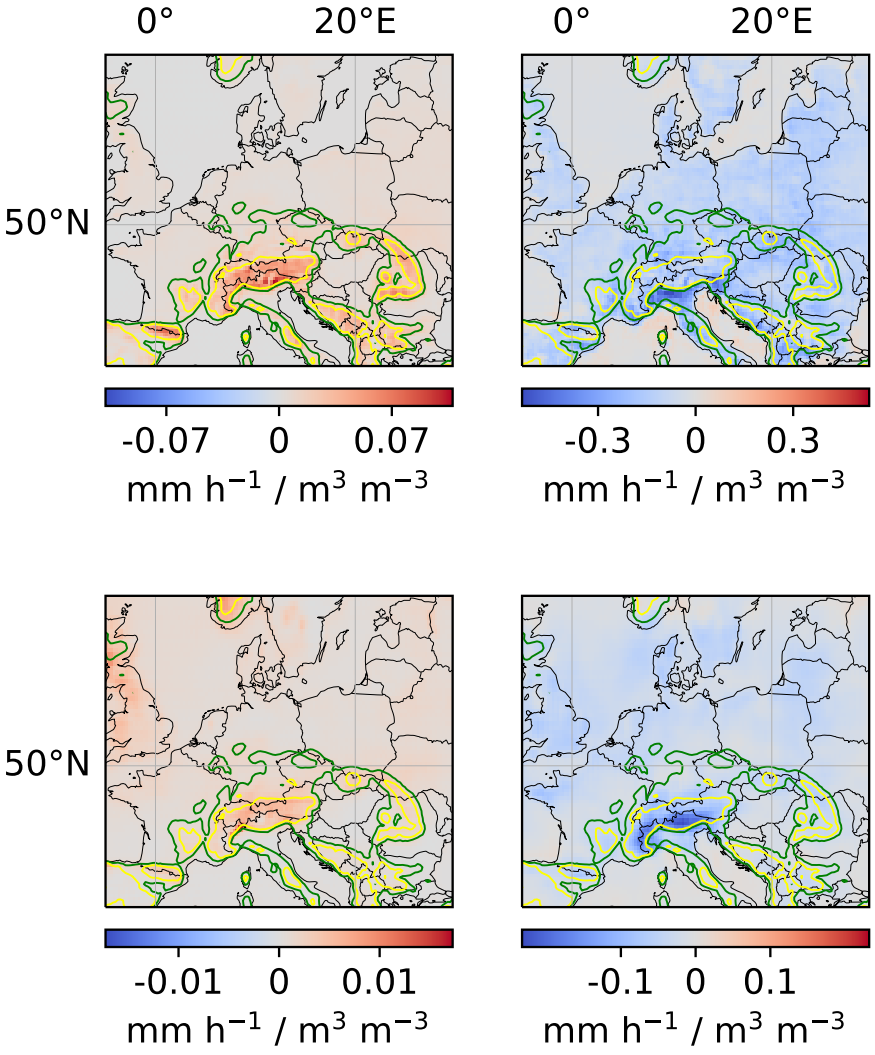

Figure A6. Local and regional soil moisture-convective precipitation and soil moisture-large-scale precipitation couplings. Left column: local ~~coupling~~couplings. Right column: regional ~~coupling~~couplings. Upper row: soil moisture-convective precipitation coupling. Lower row: soil moisture-large-scale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

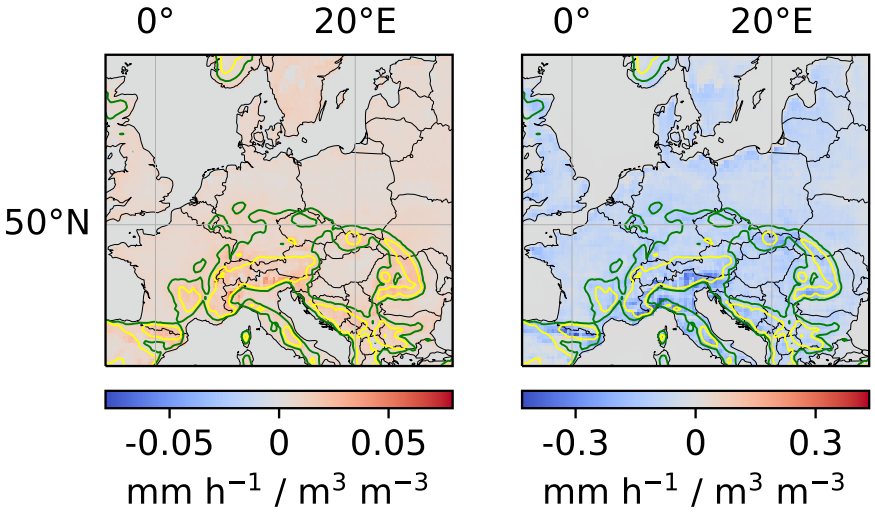

Figure A7. ~~**Product of local soil moisture-evaporation and local/ regional evaporation-precipitation coupling**~~**Product of local soil moisture-evaporation and local/ regional evaporation-precipitation couplings**. Left: product of local soil moisture-evaporation and local evaporation-precipitation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

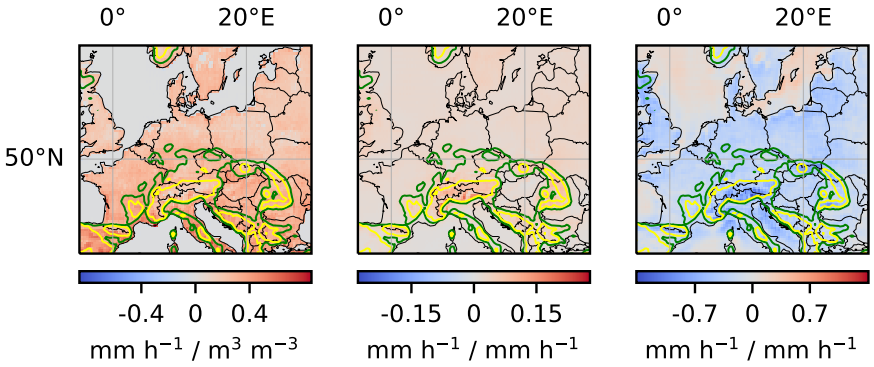

Figure A8. ~~Local soil moisture-evaporation and local and regional evaporation-precipitation coupling~~Local soil moisture-evaporation and local and regional evaporation-precipitation couplings. Left: local soil moisture-evaporation coupling. Centre: local evaporation-precipitation coupling. Right: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

Taylor, C. M., de Jeu, R. A. M., Guichard, F., Harris, P. P., and Dorigo, W. A.: Afternoon rain more likely over drier soils, Nature, 489, 423–426, https://doi.org/10.1038/nature11377, 2012.

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.