Interpretable deep learning for studying the Earth system

Soil-moisture–precipitation coupling across Europe

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Jan Tobias Tesch

aus

Troisdorf

Bonn 2023

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Stefan Kollet
2. Gutachter: Prof. Dr. Jochen Garcke

Tag der Promotion: 26.09.2023
Erscheinungsjahr: 2023

Information on the assistance received and resources used

The submitted thesis is my own work and was prepared without unauthorized assistance by others. I indicated all sources and resources used in accordance with good scientific practice. I did not use any own paper that has already been used for examination purposes.

Acknowledgements

First of all, I thank my supervisor Professor Stefan Kollet for his support, fruitful discussions and numerous helpful suggestions. Thank you very much for taking so much time.

I am also very grateful to my second supervisor Professor Jochen Garcke for his support. You accompanied me during my bachelor thesis, my master thesis and now during my doctoral thesis. Thank you very much.

Moreover, I thank the other members of my doctoral committee, Dr. Petra Friederichs and Professor Christopher McCool.

I want to thank everyone in Stefan's group as well as the other employees of IBG-3 for their helpfulness and the friendly working atmosphere. In particular, I want to thank Carl Hartick, Yueling Ma and Klaus Goergen for answering countless questions.

Special thanks go to my wife Rebekka. During the COVID-19 lockdowns, we worked at the same desk for months. You were always open for discussions and kept up my spirits.

Abstract

The Earth system is a highly complex dynamical system. While considerable process understanding has been achieved in past research, many processes and relations in the Earth system remain poorly understood due to this complexity. A better understanding of these processes and relations can improve weather and climate predictions and eventually help make decisions that protect life and property. In this thesis, I evolve the recently proposed approach of using interpretable deep learning to gain new scientific insights into the Earth system. In the approach, a deep learning model is trained to predict one Earth system variable (referred to as *target* variable) given some others as input. After training the model, the relations between input and target variables that the model learned are analyzed to gain new scientific insights. The major challenge to the approach is that the model may learn spurious correlations rather than actual causal relations. This is a challenge, not only because the scientist cannot gain new scientific insights from a model that learned spurious correlations, but also because detecting whether a given model learned spurious or causal relations is difficult in complex systems.

Here, I propose a variant approach to identify spurious correlations that any given statistical model learned. Furthermore, I develop a methodology of causal deep learning models, which combines the approach of using interpretable deep learning to gain new scientific insights with findings from causality research to actually obtain a causal deep learning model, i.e. a model that learns the *causal* relations between input and target variables. Applied to several examples from hydrometeorology, the variant approach is superior to other commonly applied approaches for identifying spurious correlations that statistical models learn. Moreover, results obtained with causal deep learning models differ entirely from results obtained with a simple linear correlation analysis, which stresses the importance of considering non-linear effects and the difference between correlation and causation.

Finally, I apply both methodologies to gain new insights into soil-moisture–precipitation coupling, i.e. the question how soil moisture affects precipitation. Improving our understanding of soil-moisture–precipitation coupling can help to better understand and mitigate extreme events like droughts and floods, and the effects of land management and climate change. The developed methodology of causal deep learning models overcomes several common limitations of previous studies on soil-moisture–precipitation coupling and reveals that an increase in local soil moisture leads to a subsequent increase in precipitation locally, and a simultaneous decrease in precipitation in a surrounding area. The non-local coupling strength exceeds the local coupling strength. These findings contribute to our understanding of soil-moisture–precipitation coupling and stress the importance of non-local effects, which have commonly been neglected in previous studies.

Contents

1. Motivation and outline **1**

2. Deep learning **7**

2.1. Interpretable deep learning 10

2.2. Physics-informed deep learning 11

2.3. U-net architecture 11

3. Learning causal relations from observations **13**

3.1. Causal discovery 16

3.2. Causal inference 17

3.3. Causal representation learning 19

4. Soil-moisture–precipitation coupling **21**

4.1. Physical processes 23

4.2. Modelling and statistical studies on soil-moisture–precipitation coupling 24

4.3. Limitations of existing approaches 25

4.4. Results from previous studies on soil-moisture–precipitation coupling 26

4.4.1. Modelling approaches 26

4.4.2. Statistical approaches 30

5. Variant approach for identifying spurious relations that deep learning models learn **35**

6. Causal deep learning models for studying the Earth system **39**

7. Converse local and non-local soil-moisture–precipitation couplings across Europe **43**

8. Conclusion and outlook **47**

8.1. Summary 49

8.2. Challenges and limitations 50

8.3. Recommendations for future work 51

References **55**

Appendix **71**

A. Variant approach for identifying spurious relations that deep learning models learn **73**

A.1. Research article 75

A.2. Supporting information 93

B. Causal deep learning models for studying the Earth system **97**

C. Converse local and non-local soil-moisture–precipitation couplings across Europe 117
C.1. Research article 119
C.2. Supporting information 139

List of Figures 155

1. Motivation and outline

The Earth system comprises countless complex processes relating Earth system variables across various spatio-temporal scales (e.g. Brutsaert, 2005; Kraus, 2004, see Figure 1.1 for processes in the water cycle as an example). Despite decades of research that has yielded valuable insights into the Earth system, many processes and relations remain poorly understood. A better understanding of these processes and relations will improve weather and climate predictions and help make decisions that protect life and property (Santanello et al., 2018). One of these poorly understood processes is soil-moisture–precipitation (SM–P) coupling, i.e. the question how soil moisture affects precipitation, which is studied in this thesis using a newly developed statistical methodology based on interpretable deep learning (DL) and causality research.

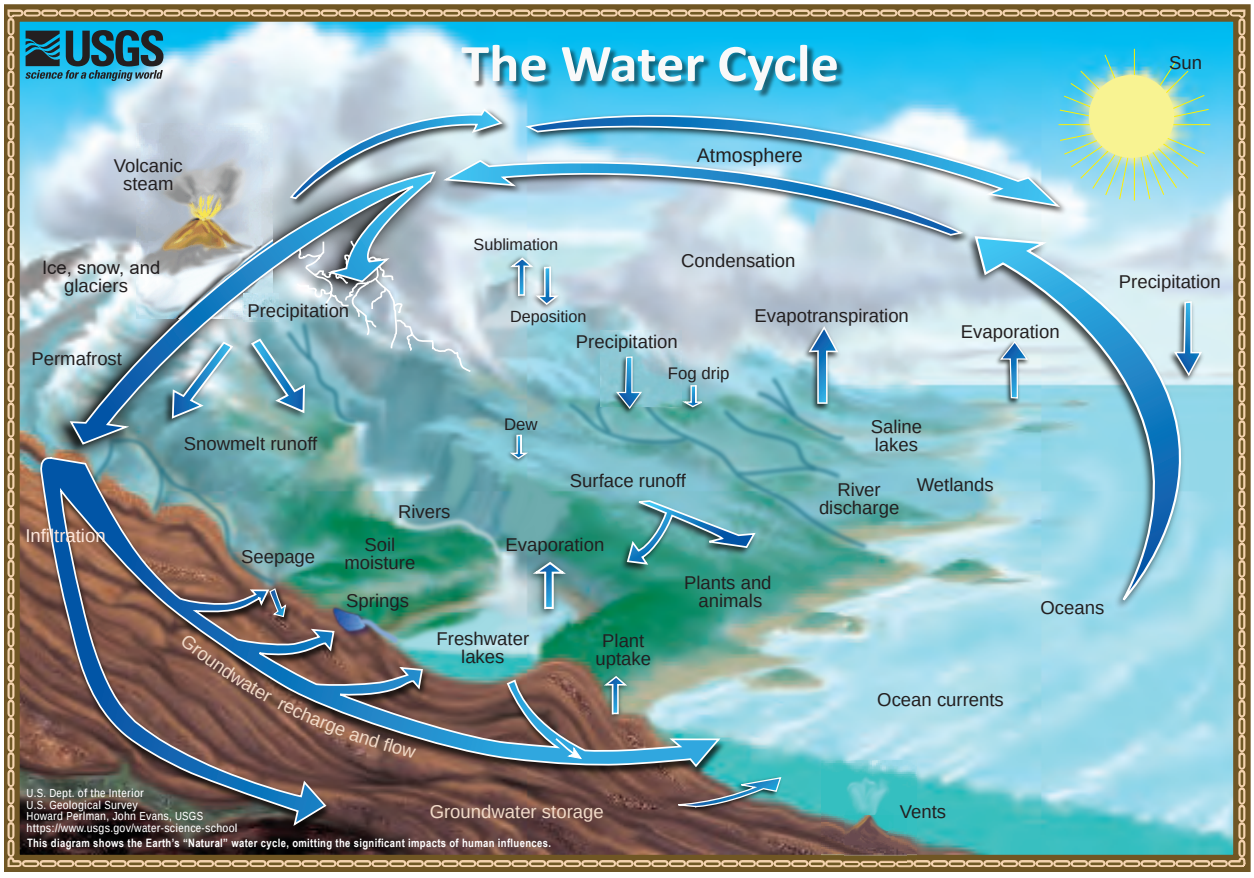

Figure 1.1: The water cycle comprises numerous complex processes linking water within and between the atmosphere, land surface, and subsurface. Source: U.S. Geological Survey's Water Science School (2019).

In light of the huge amount of available geospatial data from observations (e.g. remote sensing and in situ observations) and model simulations, statistical methodologies are increasingly used to gain new scientific insights into the Earth system (Reichstein et al., 2019). However, classical statistical methodologies have several common limitations that constrain their applicability for scientific discovery in the geosciences, e.g. requiring assumptions on linearity or locality of considered relations and hand-designed input features. DL can overcome many of these limitations. DL models are showing ongoing successes across many scientific disciplines in terms of predictive

performance (Reichstein et al., 2019; Shen, 2018). They can learn the complex, non-linear relations between Earth system variables from raw geospatial data, and the recently evolving branch of interpretable DL (Molnar, 2022; Montavon et al., 2018; Samek et al., 2021; Zhang and Zhu, 2018) allows to visualize and analyze the learned relations. In particular, training a DL model to predict one variable given some other variables and analyzing the relations that the model learned constitutes a recently suggested, promising methodology for gaining new scientific insights into the Earth system (Gagne II et al., 2019; Ham et al., 2019; McGovern et al., 2019; Roscher et al., 2020; Toms et al., 2020).

However, there is a major challenge to this approach: like every other statistical approach for studying the Earth system, it merely provides insights into statistical associations rather than actual *causal* relations between Earth system variables. Determining whether relations that a DL model learned reflect mere statistical associations (referred to as *spurious correlations*) or actual causal relations is challenging. In this thesis, I address this challenge by developing a novel methodology to identify spurious correlations. Furthermore, I combine the described approach of using interpretable DL to gain new scientific insights with a result from causality research (Pearl, 2009) stating that a statistical model may learn the actual causal impact of an input variable on a target variable if suitable additional input variables are chosen. Figure 1.2 illustrates these contributions. In the geosciences, the difference between causality and correlation is still mostly ignored (Runge et al., 2019).

I apply the proposed methodologies to study the impact of soil moisture changes on subsequent precipitation. Although known to be important for precipitation prediction, SM–P coupling remains poorly understood and an active area of research. In this thesis, it is studied across Europe at a sub-daily time scale using two different data sets, namely ERA5 climate reanalysis data (Hersbach et al., 2018) (which is deemed to be close to observations) and data from a high-resolution, convection-permitting simulation (which is deemed to better resolve the process of convection, which is essential for SM–P coupling). The developed methodologies provide new scientific insights into SM–P coupling, in particular into the importance of non-local effects.

In Chapters 2, 3 and 4 of this thesis, I provide brief introductions into DL, causality and SM–P coupling, respectively, and put this thesis into the context of previous studies in the respective areas. In Chapter 5, I describe the variant approach, which I developed to identify spurious correlations that a DL model learned. In Chapter 6, I combine the described approach of using interpretable DL to gain new scientific insights with the above-mentioned findings from causality research. Further, I illustrate the resulting methodology of causal DL models using the example of SM–P coupling in ERA5 data across Europe. Subsequently, in Chapter 7, I apply the methodology to gain new scientific insights into SM–P coupling. In particular, I compare the results obtained when applying

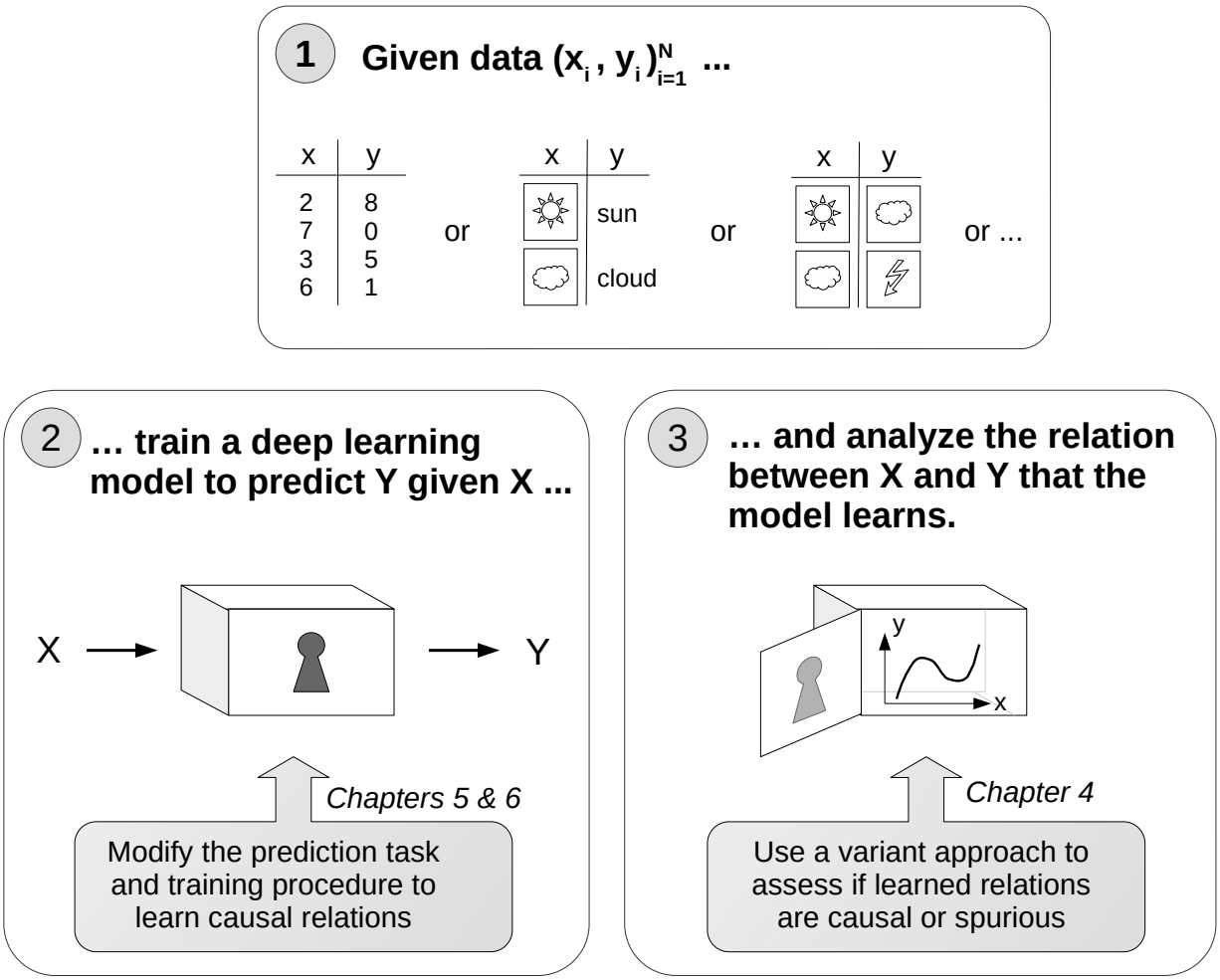

Figure 1.2: Illustration of the original methodology for obtaining new scientific insights into the Earth system and the methodological contributions in this thesis.

the methodology to ERA5 data and to data from a high-resolution, convection-permitting simulation, respectively, and compare the results to those from previous studies on SM–P coupling. Further, I provide computational details on a faster implementation of the methodology. Finally, Chapter 8 concludes this thesis by summarizing the main findings and challenges, and providing recommendations for future research.

2. Deep learning

Deep learning (DL) is a field of machine learning that is based on stacking multiple modules (so-called layers) on top of each other (Deng and Yu, 2014; Goodfellow et al., 2016; LeCun et al., 2015; Reichstein et al., 2019). Each of these layers performs simple but non-linear mathematical operations on its respective inputs (which usually constitute the outputs of the previous layer). Starting from raw inputs, each layer extracts more complex concepts (*features*), e.g. the first layer detecting the existence and orientation of edges in an input image, the second layer detecting specific arrangements of edges, and so on. Finally, the last layers map the extracted, complex concepts to some prediction quantity of interest, e.g. whether the image shows a cat or a dog (see Figure 2.1).

What sets DL apart from many classical statistical approaches is that the extracted features are not hand-designed by a human expert, but learned from raw data by minimizing some loss function, usually using backpropagation (LeCun et al., 2012) and variants of stochastic gradient descent. Combining enough simple layers, DL models can represent any function with arbitrary precision (Cybenko, 1989; Hornik, 1991; Leshno et al., 1993). Driven by breakthrough performances in image processing, in particular in the ImageNet competition in 2012, where DL models almost halved the error rates of competing image recognition approaches (Krizhevsky et al., 2012), and further breakthrough-performances in video, speech, audio and text processing (LeCun et al., 2015), DL has found its way into sciences and is showing ongoing successes across many scientific disciplines (Reichstein et al., 2019; Shen, 2018). For a detailed introduction to DL, I refer to (Goodfellow et al., 2016).

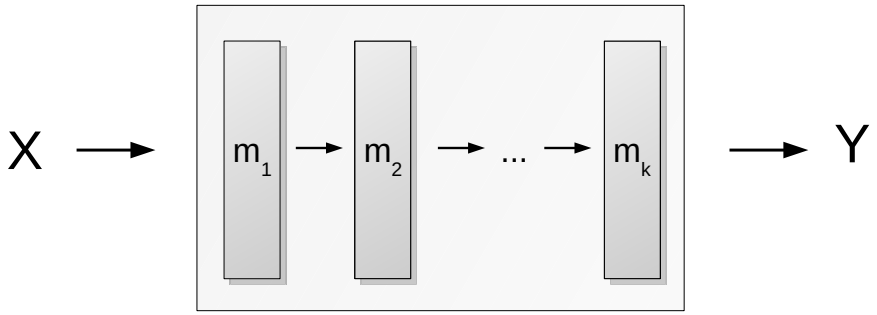

Figure 2.1: Schematic representation of a deep learning model. The input X is passed through several consecutive modules m_i (layers), which perform simple but non-linear mathematical operations on their respective inputs, to produce an output Y. For an example of a state-of-the-art DL model see Chapter 2.3 and Figure 2.2.

In the geosciences, there have been many successful applications of DL in recent years, both in research projects as well as in operational products (Camps-Valls et al., 2020). They benefit from the ever-increasing amount of geoscientific data, data from observations (e.g. remote sensing and in situ observations) and model simulations (Reichstein et al., 2019). Further, they benefit from the improved capabilities of DL to incorporate spatial and temporal structures in geoscientific data

compared to classical statistical models that require hand-designed input features. For example, detecting hurricanes in meteorological data (like fields of precipitation, meridional wind, humidity and other variables) requires to take into account spatio-temporal context and interactions between different variables. It is difficult to hand-design suitable features for such a task, which gives DL models an advantage over classical approaches (Liu et al., 2016; Racah et al., 2017; Reichstein et al., 2019).

For a textbook and reviews on DL applications in the geosciences, I refer to (Camps-Valls et al., 2021) and (Reichstein et al., 2019; Shen, 2018), respectively. Successful applications of DL in the geosciences include weather forecasting (Espeholt et al., 2022), extreme weather detection (Liu et al., 2016; Racah et al., 2017), El Niño-Southern Oscillation prediction (Ham et al., 2019), rainfall-runoff modelling (Kratzert et al., 2018), land use and land cover classification (Zhu et al., 2017), wildfire (Lee et al., 2017) and landslide detection (Liu and Wu, 2016).

2.1. Interpretable deep learning

A downside of stacking many layers and automatically learning features from data, is a potential loss in interpretability, i.e. it is less clear how inputs relate to predictions of the models than for simpler statistical models. Under the term of interpretable DL, several interpretation methods for DL models have been developed in the last years (Molnar, 2022; Montavon et al., 2018; Samek et al., 2021; Zhang and Zhu, 2018). A prominent subclass of these methods are feature importance methods, which indicate for each raw input feature (e.g. each pixel of an image) how it contributed to the prediction of the model. Among the most prominent examples for such methods are the gradients of the DL model, also called saliency maps (Simonyan et al., 2013) (see Section 2.3.5 of Appendix A.1), Layerwise Relevance Propagation (LRP; Bach et al., 2015) and Gradient-weighted Class Activation Mapping (Grad-CAM; Selvaraju et al., 2017).

On the one hand, and particularly relevant for the methodology of causal DL models developed in Chapter 6, using interpretation methods to understand a DL model's predictions can lead to new scientific insights (Gagne II et al., 2019; Ham et al., 2019; McGovern et al., 2019; Montavon et al., 2018; Roscher et al., 2020; Schütt et al., 2017; Toms et al., 2020). Ham et al. (2019), for example, identified a previously unreported precursor of the Central-Pacific El Niño type using these interpretation methods. Nevertheless, telling whether the explanations obtained with these interpretation methods reflect causal relations or spurious correlations between input and target variables is challenging even for experts. This limits the usefulness of interpretable DL for gaining new scientific insights. To tackle this challenge, I propose a variant approach, described in Chapter 5. Further, in Chapter 6, I combine the idea of using interpretable DL to gain new scientific insights with insights from causality research in order to achieve that the DL model actually learns

causal relations rather than mere statistical associations (see Figure 1.2).

On the other hand, using interpretation methods to understand why a DL model predicts what it predicts can build trust in the model (Ribeiro et al., 2016), reveal a model's limitations (Lapuschkin et al., 2019), and help improving a model (Schramowski et al., 2020). Lapuschkin et al. (2019), for example, used interpretation methods to reveal undesired behavior of several DL models, such as the reliance of an image classification model on copyright tags on certain images. Schramowski et al. (2020) detected undesired behavior of their DL model using a feature importance method and corrected the behavior by penalizing meaningless feature importance scores during the training procedure. A difficulty for using interpretation methods to detect undesired behavior of a DL model arises when the scientist cannot judge whether an obtained explanation reflects undesired behavior or not, e.g. when the relations between input and target variables are complex or unknown. The variant approach proposed in Chapter 5 allows to use interpretation methods for building trust in a model or revealing a model's limitations even in these cases.

2.2. Physics-informed deep learning

Next to interpretable DL for scientific discovery, the methodology of causal DL models developed in Chapter 6 also relates to the research branch of physics-informed DL (Kashinath et al., 2021; von Rueden et al., 2021). Physics-informed DL aims to include physical knowledge into the formulation of DL tasks in order to increase the physical consistency and performance of DL models. Examples for physics-informed DL are the inclusion of a physically motivated loss term in the training procedure (Daw et al., 2017) or physically motivated transformations of the input or target variables (Dramsch et al., 2019). The combination of interpretable DL and causality research described in Chapter 6 relates to physics-informed DL in that the choice of input variables in the approach requires certain physical knowledge of the system, in particular of its causal structure. However, in general, the motivation for including physical knowledge in physics-informed DL is to improve the performance of a DL model on a given DL task, while the motivation for including physical knowledge in the approach of causal DL models is to remove bias in the estimation of causal effects.

2.3. U-net architecture

In this thesis, I use convolutional neural networks (CNNs), a class of DL models that is particularly useful when the input in the prediction task has a grid-like topology (Goodfellow et al., 2016), e.g. one-dimensional time series of measurements at a regular time interval, or two-dimensional grids of pixels in an image. In Section 2.3.4 of Appendix A.1, I give a short introduction to the mathematical operations in CNNs. For a more in depth introduction, I refer to (Goodfellow et al.,

2016) and the countless blogs on deep learning (e.g. Amidi and Amidi).

In Chapters 6 and 7, I use a U-net architecture (Ronneberger et al., 2015, see Figure 2.2), a popular CNN architecture used when both input and predicted variables have a grid-like topology. U-nets are commonly used for image segmentation (e.g. medical image segmentation (Ronneberger et al., 2015; Siddique et al., 2021), road extraction from aerial images (Zhang et al., 2018), land cover segmentation (Rakhlin et al., 2018; Solórzano et al., 2021), and cloud detection in satellite imagery (Guo et al., 2020)) and regression tasks (e.g. precipitation prediction (Agrawal et al., 2019; Han et al., 2022; Sadeghi et al., 2020)). Alternative model architectures used when both input and predicted variables have a grid-like topology include SegNet (Badrinarayanan et al., 2017) and FCN (Long et al., 2015). In early experiments in the scope of this thesis, these alternatives as well as slight architectural variations of Figure 2.2 yielded similar sensitivities but with slightly reduced predictive performance, which is why I use the U-net architecture in this thesis. Larraondo et al. (2019) also found the U-net architecture to perform better in precipitation prediction than SegNet and FCN.

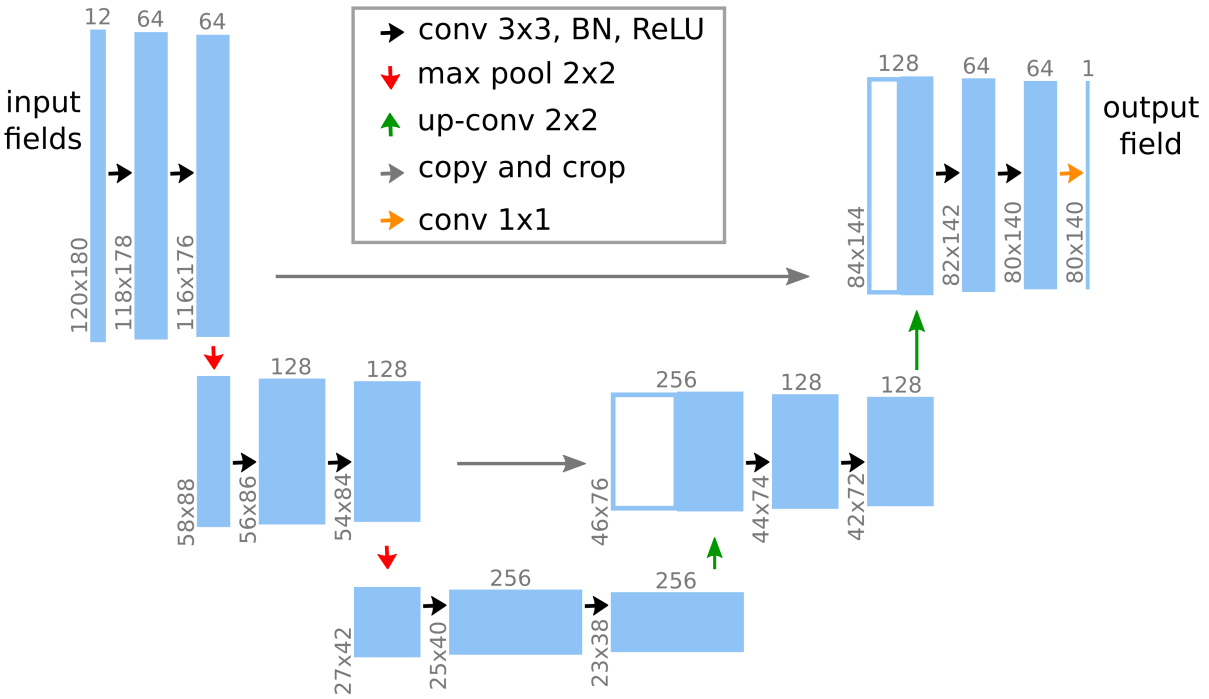

Figure 2.2: U-net model architecture. The input to the model is represented by the leftmost blue box and, in this example, consists of 12 variables at 120×180 input pixels. It is passed through multiple sequential layers represented by the arrows. Performing simple mathematical operations on its respective inputs, each layer produces an output represented by the next blue box. In general, this output differs in shape from the input, as indicated by the grey upright and rotated numbers. This output is fed to the next module until the rightmost blue box represents the output of the model, in this example a prediction at 80×140 target pixels. For details on the mathematical operations, I refer to (Amidi and Amidi; Goodfellow et al., 2016; Ronneberger et al., 2015). Figure originally published in (Tesch et al., 2023a).

3. Learning causal relations from observations

Understanding causes and effects allows to make sense of and systematically influence the world around us. Therefore, one of the key objectives in science is to identify and estimate causal relations. A standard class of approaches for studying causal relations is based on experiments that intervene into the system of interest and evaluate the effects of these interventions (Runge et al., 2019). Randomized controlled trials (RCTs), for instance, are a standard approach for studying causal relations in medicine and the social sciences (Imbens and Rubin, 2015; Sibbald and Roland, 1998). For example, for studying the impact of food supplements on mortality, participants in a study may be randomly partitioned into participants receiving supplements (*intervention group*) and participants not receiving supplements (*control group*) and the difference in mortality between intervention and control groups assessed (Autier and Gandini, 2007). However, in many cases, conducting experiments is either infeasible or ethically problematic. For example, we should not conduct large-scale experiments on the Earth's atmosphere (Runge et al., 2019). Further, while RCTs can yield average causal effects, e.g. the difference between average mortality in intervention and control groups, often, they cannot yield individual causal effects, because only one outcome (e.g. either with *or* without treatment) is observed for each instance (Guo et al., 2021; Knaus et al., 2020; Yoon et al., 2018). In some sciences, including the geosciences, both issues can be avoided by resorting to numerical simulations, e.g. simulating precipitation on a certain day with different initial soil moisture conditions to evaluate the causal effect of soil moisture on precipitation on that specific day. However, numerical simulations bring their own challenges like high (computational) costs and strong assumptions on the system (Runge et al., 2019).

Another class of approaches for studying causal relations, which is adopted in this thesis, is to learn from purely observational data of the system of interest (Guo et al., 2021; Runge et al., 2019, see Figure 3.1). However, there is a major challenge for these approaches, namely the difference between statistical associations and causal relations. Indeed, according to Reichenbach's common cause principle (Reichenbach, 1956), a statistical association between two variables X and Y implies that there exists a variable Z that causally influences both (where Z might also be X or Y as special cases). However, a statistical association between X and Y is far from implying a causal relation between X and Y. For example, the frequency of storks is correlated with human birth rates (Matthews, 2000; Schölkopf et al., 2021). However, the correlation between the frequency of storks and human birth rates is due to common causes (e.g. economic development) rather than due to a causal link between them.

While most statistical concepts, e.g. conditional expectations, are fully definable in terms of the joint probability distribution of observed variables, studying causality from observational data requires to introduce new notation for expressing causal relations. Further, it requires a priori causal assumptions that are not (fully) testable in observational studies (e.g. that no unknown or unobserved

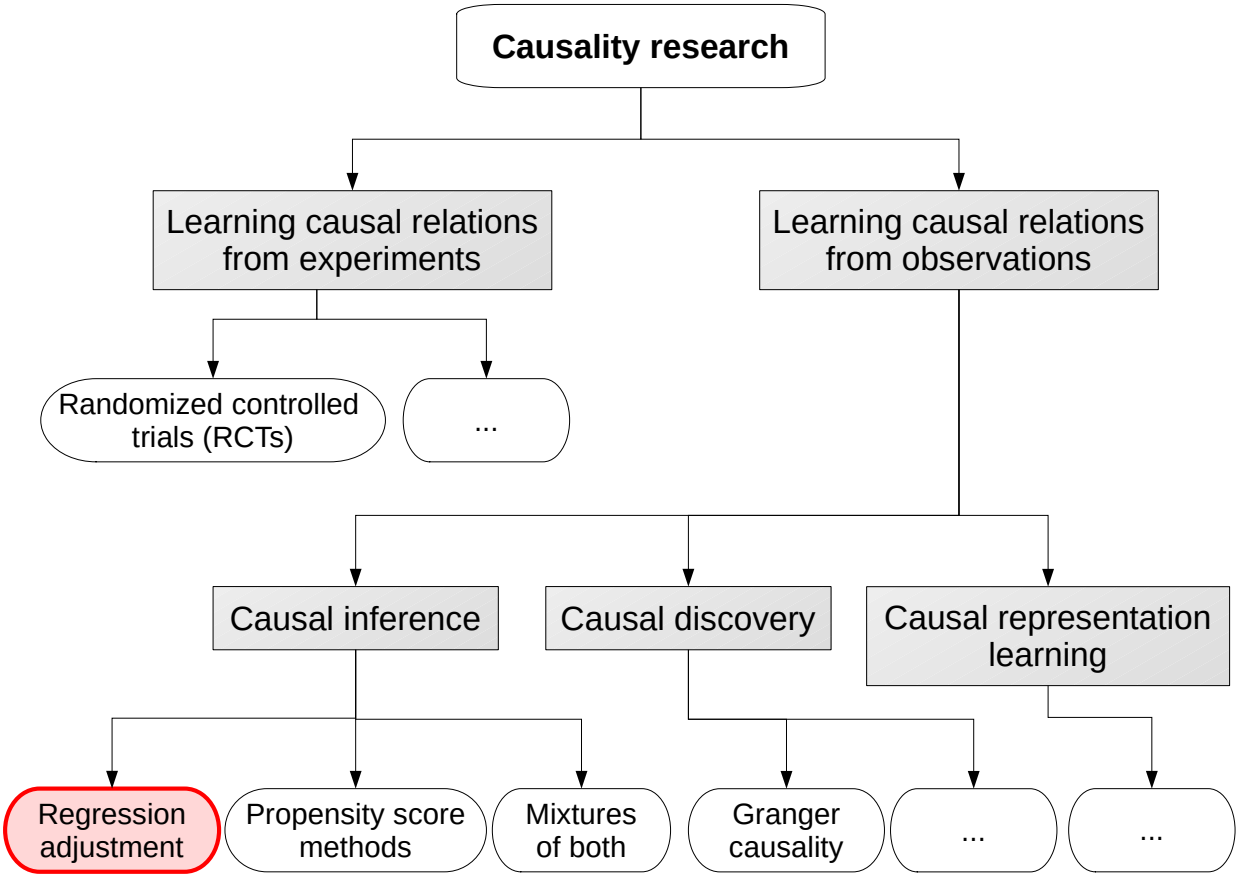

Figure 3.1: Overview over different areas (boxes) and methodologies (ellipses) of causality research. In this thesis, regression adjustment is used.

confounders, variables that may affect the considered causal relations, exist) (Pearl, 2009). A general theory for causality is described in (Pearl, 2009). It allows to represent causal questions in mathematical language and to systematically determine what assumptions or measurements are necessary to answer these questions. The theory is based on structural causal models (SCMs). These consist of two components, a causal graph and structural equations, which encode the causal structure of a system (see Figure 3.2 for a simple example). An introduction to this theory and SCMs is given in Section 2.1 of Appendix B and, for example, in (Guo et al., 2021; Massmann et al., 2021; Pearl, 2009).

3.1. Causal discovery

One subclass of approaches for studying causal relations from observational data focuses on causal discovery. Given a set of observed variables $\{X_i\}_{i=1}^n$, approaches in this subclass aim to determine for each pair (X_i, X_j), $i \neq j$, whether variable X_i changes if variable X_j is modified. More formally, in the framework of SCMs, they aim to identify the causal graph underlying the observed variables $\{X_i\}_{i=1}^n$ (Guo et al., 2021; Massmann et al., 2021; Runge et al., 2019).

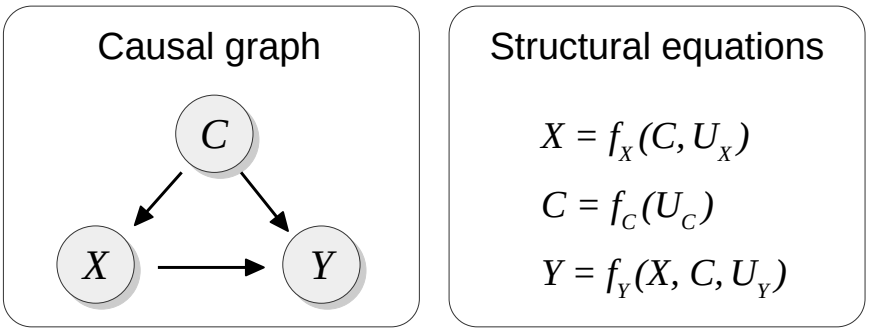

Figure 3.2: A simple structural causal model consisting of three variables, X, Y, and C. The arrows in the causal graph encode the causal dependencies, e.g. X causally affects Y and is causally affected by C. The structural equations describe these dependencies. f_X, f_C and f_Y are functions, and U_X, U_C and U_Y are random variables representing potential chaos and variables not included in the causal graph explicitly.

A prominent example of an approach for causal discovery is Granger causality (Granger, 1969; Papagiannopoulou et al., 2017; Runge et al., 2019). The original idea of this approach is to test whether the error of a model predicting the current value of a time series Y from its own and further covariates' past decreases when the past of a time series X is included additionally. If this is the case, X Granger-causes Y. There are several works using Granger causality in the geosciences, e.g. Papagiannopoulou et al. (2017) investigate climate-vegetation dynamics using Granger causality.

Another approach for causal discovery is the PC algorithm (Spirtes et al., 2000), a variant of which is applied for example in (Barnes et al., 2019; Ebert-Uphoff and Deng, 2012, 2017) to study atmospheric teleconnections. Given a set of observed variables $\{X_i\}_{i=1}^n$, the algorithm starts with edges between every pair of variables (X_i, X_j), $i \neq j$. Next, it evaluates for each pair of variables whether X_i is conditionally independent of X_j given any subset of the remaining variables. If this is the case, the edge between the variables is removed, because there cannot be a direct causal link between these variables. Edges that remain upon termination of the algorithm represent potential causal links between the associated variables. Further applications of causal methods in the geosciences are described in (Runge et al., 2019). Note that so far, in the geosciences, most works on causality focus on causal discovery.

3.2. Causal inference

In this thesis, I focus on causal inference (also called causal estimation) rather than causal discovery. Approaches for causal inference focus on determining the strength of causal relations, i.e. on answering *how much* a specific variable would change if another variable was modified (Guo et al., 2021; Massmann et al., 2021; Pearl, 2009). In general, they assume that the basic causal structure (i.e. the causal graph in the framework of SCMs) is known and the causal effect of interest is identifiable (Pearl, 2009). A causal effect might for example not be identifiable if there are

unobserved confounders that cannot be adjusted for (see below). Approaches for causal inference may be divided into approaches for cases with and without unobserved variables, respectively. In this thesis, I focus on the case that the scientist has access to all relevant variables, i.e. that there are no unobserved variables. This is reasonable when estimating causal effects from reanalysis data or output data of a numerical simulation, where the output comprises all relevant variables, but has to be reconsidered when aiming for direct causal inference from observational data of the Earth system.

Most approaches for estimating the causal effect of a variable X on Y from data without unobserved variables are either based on regression adjustment or propensity scores (Guo et al., 2021). In regression adjustment, a statistical model is trained to predict Y given X and further input (*adjustment*) variables $C_i, i = 1, \dots, k$. To obtain an unbiased estimate of the causal effect of X on Y, the input variables C_i have to form an admissible (also called "sufficient") set, either fulfilling the backdoor criterion (Massmann et al., 2021; Pearl, 2009), or the slightly more general criterion from (Perković et al., 2018) adopted in Chapters 6 and 7. So far, like causal inference in general, regression adjustment has received very little attention in the geosciences (Kretschmer et al., 2016; Massmann et al., 2021; Runge et al., 2014). Note that there is some (theoretical) work on how to optimally choose the input variables C_i whenever there exist more than one admissible set (Henckel et al., 2022; Perković et al., 2018; Rotnitzky and Smucler, 2020; Runge, 2021; Witte et al., 2020), which however does not apply to the case of highly complex non-linear systems and the DL models considered in this work.

The term *propensity score* refers to the value $\mathbb{P}(X = x|\{C_i = c_i\}_{i=1}^k)$, i.e. the probability of observing a value X of x given values c_i of the covariates C_i (Guo et al., 2021). To motivate propensity scores, consider the example of estimating the causal effect of a binary treatment X (e.g. food supplements yes or no) on some outcome variable (e.g. mortality). In RCTs, the causal effect of the treatment could be determined by simply considering the difference between the average outcome for all treated individuals and the average outcome for all untreated individuals, because treatment was assigned randomly, i.e. *the probability of being treated or not was identical for each individual*. Given propensity scores, groups of treated and untreated individuals can be built, where, as in RCTs, all individuals have the same probability of being treated or not (the same propensity score). For each of these groups, the causal effect of actually receiving the treatment can be determined by simply building the difference between the average outcome for all treated individuals and the average outcome for all untreated individuals as in RCTs. This method is referred to as propensity score matching or propensity score stratification (Guo et al., 2021). Other causal inference methods based on propensity scores use the propensity scores to weight samples according to their inverse propensity score to "synthesize a RCT", or use them in

combination with regression adjustment (Guo et al., 2021). While propensity score methods may have some theoretical advantages over simple regression adjustment in some cases (Elze et al., 2017), in practice they are not necessarily superior to regression adjustment (Cepeda, 2003; Elze et al., 2017).

Most of the works on causal inference focus on estimating the causal effect of a one-dimensional, binary treatment on some outcome variable of interest, e.g. the causal effect of food supplements yes or no on mortality. This is likely due to the binary nature of standard RCTs, which are the gold standard for studying (average) causal effects (Hariton and Locascio, 2018).

As there are also settings with continuous treatments, methods and theory for learning causality from observations have partly been extended to continuous treatments (Galagate, 2016). However, when studying continuous treatments, there is no longer a unique quantity that represents the (average) causal effect of the treatment on the outcome variable as in the case of a binary treatment. A common task when studying continuous treatments is to determine the expected outcome for setting the treatment variable to a range of possible values (independent of the observed values). In this thesis, I instead consider the question how the outcome is expected to change, if the treatment variable is slightly modified from the originally observed value (see Chapter 6). This does not seem to be a common research question, although it provides interesting insights into the Earth system (see Chapters 7 and 8).

The causal variables considered in Chapters 6 and 7 represent gridded spatial variables, e.g. soil moisture at a grid of pixels, and the considered causal graph holds locally for each pixel as well as globally for all pixels. This setting differs from other imaging settings, where causal variables have to be derived from the images first (see Chapter 3.3). To the best of my knowledge, the setting considered in this thesis has not been considered before. Therefore, and because many other approaches for estimating causal effects do not directly extend from the case of a one-dimensional, binary treatment variable to the considered case of a high-dimensional, continuous treatment variable (Hill, 2011; Knaus et al., 2020; Shi et al., 2019; Wager and Athey, 2018; Yoon et al., 2018) or require further assumptions, e.g. on the relations in the considered system (Chernozhukov et al., 2018), I use regression adjustment in this thesis. Nevertheless, adapting other approaches for estimating causal effects to the cases considered in this thesis provides an interesting avenue for future work.

3.3. Causal representation learning

Next to causal discovery and causal inference, another line of causality research, which is not relevant to this thesis, focuses on causal representation learning. This research considers un-

structured data such as images. In contrast to structured data, where each variable (e.g. in this thesis, each meteorological variable at some location) represents a variable in a causal graph, in unstructured data, causal variables have to be derived from the data first. For example, when using medical images to predict the likelihood of developing a disease, the causal variables are not the values of fixed pixels, but for instance the size of some organs or other specific patterns in the image. Within the geosciences, causal representation learning might be used to extract higher-level variables representing climatological subprocesses from gridded Earth system variables (e.g. strength of the jet stream) (Runge et al., 2019) rather than directly using the gridded Earth system variables in the causal graph as done in this thesis. For a review of causal representation learning, I refer to (Schölkopf et al., 2021).

4. Soil-moisture–precipitation coupling

Accurate precipitation prediction in weather and climate simulations can help to better understand and mitigate extreme events like droughts and floods, and the effects of land management and climate change. Soil-moisture–precipitation (SM–P) coupling refers to the impact of soil moisture on precipitation. Although this impact is known to be important, and despite decades of research, sophisticated numerical models and a plethora of observational data, SM–P coupling remains poorly understood and an active area of research. In what follows, I present the physical processes relevant to SM–P coupling, introduce the classes of methods used to study the coupling and lay out common limitations. Lastly, I review recent studies on SM–P coupling that are similar to this thesis in terms of the considered research question (i.e. how does a change in soil moisture affect precipitation) and the considered spatial and temporal scales (i.e. studies on diurnal timescales and local soil moisture changes). For comprehensive reviews on SM–P coupling, I refer to (Liu et al., 2022; Santanello et al., 2018; Seneviratne et al., 2010)

4.1. Physical processes

Soil moisture affects precipitation via its influence on the land surface water and energy balances. In particular, increased soil moisture leads to an increase in available energy at the land surface because it decreases albedo and surface temperature and thereby reduces outgoing short- and longwave radiation (Eltahir, 1998; Hauck et al., 2011; Schär et al., 1999). Moreover, increased soil moisture increases the fraction of available energy that is transformed into latent heat of evaporation, while decreasing the fraction that is transformed into sensible heat (Seneviratne et al., 2010).

These controls of soil moisture on the land surface water and energy balances give rise to a complex interplay of processes affecting precipitation (see Figure 4.1). Namely, increased latent heat flux can increase precipitation via an increase in atmospheric water content (referred to by *moisture recycling*, Eltahir, 1998) or via an increase in moist static energy in the boundary layer (Findell and Eltahir, 2003a,b; Gentine et al., 2013). However, increased sensible heat flux has been associated with stronger thermals and growth of the atmospheric boundary layer, which can also trigger precipitation (Findell and Eltahir, 2003a,b; Gentine et al., 2013; Hohenegger et al., 2009). Finally, spatial heterogeneity in sensible and latent heat fluxes can cause spatial heterogeneity in the temperature and humidity profiles of the lower atmosphere, which in turn can affect mesoscale circulations and precipitation (Adler et al., 2011; Eltahir, 1998; Gentine et al., 2019; Taylor, 2015; Taylor et al., 2011). Due to the complex interplay of these processes, increased soil moisture can lead to both increases and decreases in precipitation and the effect can arguably not be distinguished from theoretical considerations alone.

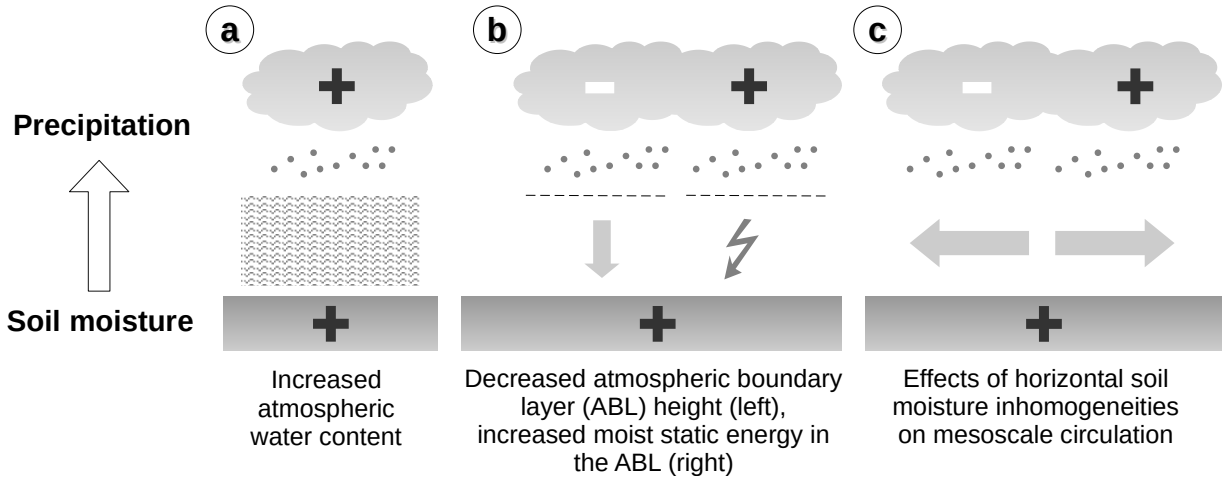

Figure 4.1: Concurring effects of soil moisture increases on subsequent precipitation. Adapted from (Tesch et al., 2023a).

4.2. Modelling and statistical studies on soil-moisture–precipitation coupling

Over the last decades, many studies have investigated SM–P coupling using different modelling and statistical approaches. By modelling approaches, I refer to approaches that study SM–P coupling by performing multiple simulations with varying soil moisture conditions, e.g. with varying soil moisture initializations. In contrast, by statistical approaches, I refer to analyses of observational data, reanalysis data and analyses of other simulations. From a causality perspective, modelling approaches correspond to experiments that intervene into the system of interest and evaluate the effects of these interventions. Accordingly, statistical approaches correspond to approaches that learn causal relations from (observational) data (see Chapter 3).

Likely the most prominent example for a modelling approach is the Global Land Atmosphere Coupling Experiment (GLACE; Guo et al., 2006; Koster et al., 2004). In this experiment, twelve atmospheric general circulation models (AGCMs) were used to identify hot spots of SM–P coupling, i.e. specific locations where soil moisture variations have a substantial impact on precipitation. Each AGCM was used to generate an ensemble of 16 simulations of boreal summer (June through August) in which soil moisture varied between the simulations, and an ensemble of 16 simulations in which soil moisture was taken from one of the former simulations and forced to be the same across the ensemble. Then, hot spots of SM–P coupling were identified by identifying regions where the difference in the variability of precipitation between the first 16 and the latter 16 simulations was largest. It was found that the major hot spots lie in transition zones between wet and dry climates, where evaporation is very sensitive to soil moisture and high enough to have a substantial impact on precipitation.

A prominent example for a statistical approach for studying SM–P coupling is (Taylor et al., 2012). First, the authors defined that a rain event at a location occurred if afternoon rain at the location exceeded 3 mm and was larger than afternoon rain at locations in a specified surrounding area. For each such rain event, they computed the difference between the early morning soil moisture anomaly at the rain location and at the location in the surrounding area with the least afternoon rain. Next, they evaluated whether, on average, this difference was larger or smaller than the difference on non-event days. In an attempt to prevent confounding, they excluded days with more than 1 mm precipitation in the morning and mountainous as well as coastal regions from their analysis. Applying the methodology to different observational data sets, they concluded that there is a preference for afternoon precipitation over drier soils, whereas they found a preference for afternoon precipitation over wetter soils when applying the methodology to output from various global simulation models.

Next to these classes of approaches, there are also approaches based on water-vapour tagging in climate models or computation of back-trajectories (Seneviratne et al., 2010). I do not discuss these approaches here, because they consider only a particular aspect of soil-moisture–precipitation coupling, namely moisture recycling (Figure 4.1a).

4.3. Limitations of existing approaches

Although many different approaches have been applied for studying SM–P coupling, there are several common limitations. Modelling approaches, on the one hand, have high computational costs and, even more importantly, rely on the correct representation of SM–P coupling in the considered models. However, there exist countless uncertainties with respect to Earth system models. These uncertainties are for example demonstrated by a high variability in SM–P coupling between the AGCMs in the GLACE study described above (Koster et al., 2004) and by the opposing signs of SM–P coupling that (Taylor et al., 2012) found when applying the above-described statistical methodology to observational and modelling data sets, respectively.

One option for reducing these uncertainties is to perform simulations at higher resolution. This allows to represent crucial processes like convection and thermally driven circulations more realistically than in standard, lower-resolution Earth system models (Hohenegger et al., 2009; Leutwyler et al., 2021). While computational advances in the last years have enabled the use of high-resolution, convection-permitting Earth system models, computational constraints still limit the suite of experiments that can be performed with these models to study SM–P coupling. Indeed, (Leutwyler et al., 2021) is the only study on SM–P coupling that uses convection-permitting simulations at continental scale and spanning several years instead of smaller domains and/or considering case studies on daily time scales. Moreover, they considered only three different soil moisture

configurations at these spatial and temporal scales (realistic initial soil moisture and homogeneous perturbations of initial soil moisture saturations by ± 25 % across the entire simulation domain). In addition to approaches based on convection-permitting models still being computationally constrained, there are also many uncertainties remaining in these models with respect to precipitation (Cioni and Hohenegger, 2017; Kendon et al., 2021), as for example illustrated by the large differences in precipitation between the members of the first multi-model ensemble of regional climate simulations at kilometer-scale resolution (Ban et al., 2021).

Statistical approaches, on the other hand, usually have much lower computational costs and can directly be applied to observational and simulation data sets (observational data sets of course bringing their own uncertainties and challenges like availability and missing data (Guillod et al., 2014; Santanello et al., 2018)), relaxing the above limitations. However, current statistical approaches are also often limited, for example due to strong assumptions like linearity or locality of SM–P coupling, the negligence of the difference between correlation and causation (see Chapter 3), and metrics that are difficult to interpret. The latter limitation is for example illustrated by the methodology from (Taylor et al., 2012) described above. Considering (pixel-wise) soil moisture *anomalies* in the computed difference and defining rain events as precipitation being larger than in a specific surrounding area makes the interpretation of the results difficult, e.g. because a pixel with a dry anomaly may still be wetter than a neighboring pixel with a wet anomaly, and because the absolute values of precipitation are mostly ignored in this definition. Moreover, this metric does not seem to be suitable to answer the classical question how a change in soil moisture affects precipitation. The statistical approach of causal deep learning (DL) models developed in Chapter 6 and applied to SM–P coupling in Chapter 7 overcomes these limitations of current statistical approaches.

4.4. Results from previous studies on soil-moisture–precipitation coupling

In this section, I review recent studies on SM–P coupling that address a similar research question as Chapter 7 of this thesis (i.e. how does a change in soil moisture affect precipitation) and consider similar spatial and temporal scales (i.e. diurnal timescales and local soil moisture changes).

4.4.1. Modelling approaches

Most modelling studies based on low-resolution general circulation models indicated positive SM–P coupling, i.e. an increase in precipitation for increased soil moisture (Seneviratne et al., 2010; Taylor et al., 2012). However, it has been shown that SM–P coupling is sensitive to the parameterization of convection in low-resolution modelling frameworks to an extent that even the sign of the coupling may be reversed (Hohenegger et al., 2009; Leutwyler et al., 2021; Taylor et al.,

2013). Convection-permitting simulations have been found to agree better with observations (Hohenegger et al., 2009; Leutwyler et al., 2021; Taylor et al., 2013), although many uncertainties remain (see limitations described above). Therefore, in the following review of modelling studies, I focus on studies using convection-permitting simulations. Since all studies differ greatly in the regions considered and time periods simulated, they are listed chronologically. A summary is given afterwards.

Hohenegger et al. (2009) simulated a single month with extremely warm temperatures, weak synoptic-scale forcing, and enhanced convective activity over a 1100 km $\times$ 700 km domain covering the Alpine region. They performed simulations with realistic initial soil moisture as well as with uniformly perturbed initial soil moisture (± 30 %). At most locations, they observed a negative coupling, i.e. a decrease in accumulated precipitation for simulations with higher initial soil moisture. They explained the result by the existence of a shallow layer of stable air sitting on top of the planetary boundary layer, and shallow clouds in the dry run being more likely to transform into deep convective cells due to stronger thermals.

Hauck et al. (2011) simulated three separate 24 h case studies exhibiting different trigger mechanisms of convection initiation over a 1200 km $\times$ 1300 km domain with mountainous terrain in central Europe. They simulated each day with realistic initial soil moisture and with uniformly perturbed initial soil moisture (± 25 %), respectively. Their results showed no simple relationship regarding the sign of SM–P coupling. For instance, for one case study, both a decrease and an increase in initial soil moisture led to a decrease in precipitation. They explained this by boundary layer dynamics, arguing that an increase in soil moisture inhibited convective activity due to a lack of thermal forcing, while a decrease in soil moisture inhibited convective activity due to a decrease in convective available potential energy (CAPE).

Barthlott and Kalthoff (2011) simulated a single summer day with weak synoptic forcing over a similar domain as Hauck et al. (2011). They performed simulations with realistic initial soil moisture and with initial soil moisture varying uniformly from 50 % to 150 % of the reference simulation in steps of 5 %. They found a systematic increase in regionally averaged precipitation for increasing soil moisture in the drier than reference runs (i.e. a positive coupling). For wetter than reference runs, they found that precipitation amounts fluctuated around 80 % to 90 % of the value of the reference run. In addition, they noted that maximum precipitation in the domain increased when initial soil moisture was increased between 50 % and 125 % of the reference run. The fraction of the domain that received precipitation during the simulated day increased when initial soil moisture was increased between 50 % and 85 %. For further increases in initial soil moisture, this fraction decreased again. The non-linear behavior of total precipitation amounts, maximum precipitation in the domain and fraction of the domain receiving precipitation with respect to changes in initial soil

moisture emphasizes the complexity of SM–P coupling.

Imamovic et al. (2017) simulated multiple times a five day period with typical European summer day conditions. They considered an initially resting atmosphere over an artificial 256 km $\times$ 256 km domain with a central, Gaussian-shaped mountain with a radius of approximately 30 km. Between the simulations, they varied the height of the mountain (0-500 m) and the initial soil moisture conditions. Namely, they ran reference simulations with domain-wide initial soil moisture saturation of 60 % (typical European conditions) and additional simulations with initial soil moisture saturation varying from 70 % to 130 % of the reference run in steps of 10 %. Moreover, they ran simulations with initial soil moisture saturation varying only at the central mountain from the reference simulations. They observed a systematic increase in accumulated precipitation for the domain-wide increases in initial soil moisture (i.e. a positive coupling), while they observed a mainly local decrease in accumulated precipitation for the local increases at the central mountain (i.e. a negative coupling). The latter effect, in particular, was weaker, when the height of the central mountain was increased. They explained the increases in precipitation for domain-wide increases in soil moisture by an increase in regional moisture recycling. The decrease in precipitation for increases in soil moisture at the central mountain was explained by drier mountains strengthening mountain-valley circulations and thus increasing accumulated precipitation. They hypothesized that this effect was weaker for higher mountains because mountain-valley circulations are stronger for higher mountains and may therefore be less affected by soil moisture.

Cioni and Hohenegger (2017) simulated two separate days over an artificial 100 km $\times$ 100 km domain at even higher resolution using a large-eddy simulation model. They considered homogeneous soil moisture initializations varying from 40 % saturation to 100 %. They discovered that total precipitation was always decreased over dry soils (positive coupling) although convection can be triggered earlier over dry soils than over wet soils under certain atmospheric conditions. Further, they saw that large-scale effects or winds can reduce the strength of SM–P coupling.

Baur et al. (2018) simulated eleven separate 24 h case studies over a 1200 km $\times$ 1300 km domain in central Europe. In addition to simulations with realistic initial soil moisture conditions, they considered simulations with soil-moisture bias of ± 25 %, combined with different soil-moisture heterogeneity length-scales ranging from 30 to 140 km introduced by chessboard patterns. They discovered that precipitation averaged over the domain increased, when the initial soil moisture bias was increased (positive coupling), while precipitation tended to occur over drier soils due to thermally induced vertical circulations and background wind causing updraft regions at the downstream flank of dry patches. Further, they observed only weak SM–P coupling for the four cases exhibiting moderate rather than weak synoptic forcing.

Henneberg et al. (2018) simulated a single day with convective precipitation and strong synoptic forcing over a 400 km $\times$ 450 km domain in northern Germany. They performed several simulations with extreme, homogeneous changes in initial soil moisture (completely dry and +50 %) over the entire or parts of the domain, and realistic changes in initial soil moisture (taken from different days), respectively. To assess the uncertainty of the observed SM–P coupling, they created ensembles by slightly shifting the domain or initialization time. Only in experiments with unrealistically large variations of soil moisture, they found changes in precipitation to exceed the model spread, indicating the danger of potential impacts of soil moisture being masked by the precise model setup and the chaotic nature of convection. Besides, they observed no clear sign of SM–P coupling, but showed that both an increase and decrease in soil moisture can lead to a decrease in precipitation.

Schneider et al. (2019) simulated six separate 24 h case studies, three of them exhibiting weak and three strong synoptic forcing, over a 750 km $\times$ 700 km domain in central Europe. They performed several simulations with initial soil moisture differing homogeneously or heterogeneously (e.g. in chessboard patterns as in (Baur et al., 2018) described above) from a reference run and found that an increase in soil moisture led in most cases to an increase in domain-averaged precipitation (positive coupling), while they did not observe significant impacts of soil moisture heterogeneity. In addition, they discovered that the coupling was on average stronger for weak than for strong synoptic forcing.

Leutwyler et al. (2021) simulated ten summer seasons in continental Europe, each with realistic initial spring soil moisture and with perturbations of initial soil moisture saturations by ± 25 % in parts and the entire region, respectively. They found that a uniform increase in soil moisture led to an increase in precipitation (positive coupling), while the effect of subcontinental variations in soil moisture was more complex. They attributed the complexity of SM–P coupling to their observation that an increase in soil moisture led to less triggered convection events due to less thermal circulation, but, at the same time, to more intense events due to larger CAPE values. They identified the largest difference between wet and dry runs in the Alpine region. Because differences in evaporation between wet and dry runs were small in the Alpine region, they hypothesized that this was due to more humidity being advected to the Alpine region from neighboring regions in the wet runs.

Summarizing, modelling studies based on high-resolution simulations found positive as well as negative impacts of soil moisture on precipitation. However, on average, it appears that soil moisture changes at large scales (e.g. domain-wide) have a positive impact on regionally averaged precipitation, while the impact of soil moisture changes at smaller scales (e.g. parts of the domain) is less clear. Furthermore, soil moisture increases seem to have a negative impact on the prob-

ability of precipitation events. Lastly, stronger synoptic forcing appears to weaken SM–P coupling and convection appears to initiate more often over soils that are relatively dry compared to their surrounding areas.

4.4.2. Statistical approaches

Next to modelling studies, several studies on SM–P coupling used statistical approaches. These approaches have been applied to observational data, reanalysis data, but also modelling data. In the latter case, they differ from the above modelling studies in that they did not use the model to explicitly simulate the effect of soil moisture changes via manipulation of soil moisture conditions (see the definition of modelling studies in Chapter 4.2).

Findell et al. (2011) considered data from the North American Regional Reanalsysis (NARR; Mesinger et al., 2006) for 25 summer seasons and studied the impact of morning evaporative fraction (EF; ratio of latent heat over the sum of latent and sensible heat) on the frequency and magnitude of afternoon rainfall for each pixel (separately). In the analysis, they partitioned the data with respect to the early morning atmospheric state (using the CTP-HI$_{low}$ framework developed in (Findell and Eltahir, 2003a,b)), and ignored days with precipitation in the morning and days with atmospheric conditions that are too stable to support convection to mitigate confounding effects due to large-scale synoptic systems and precipitation persistence. They detected that high EF enhanced the probability, but only slightly the intensity of afternoon rainfall, in parts of the study region (positive EF–precipitation coupling) while not affecting afternoon rainfall in other parts. Aires et al. (2014) extended that analysis using a neural network approach rather than the simpler binning approach in (Findell et al., 2011). They confirmed that an increase in EF leads to an increase in precipitation frequency (positive EF–precipitation-probability coupling), but depending on the considered region to either an increase or a decrease in precipitation magnitude. Guillod et al. (2014) performed a similar analysis using various observational data sets in addition to NARR and found a positive EF–precipitation-probability coupling in some regions. However, they detected large differences when using different data sets due to large uncertainties in the EF data, and found that the obtained positive EF–precipitation-probability coupling might to a large extent be explained by the confounding effect of precipitation persistence.

Froidevaux et al. (2014) considered data from three, several weeks long convection-permitting simulations over an artificial region resembling a large and flat midlatitude grassland area in summer under constant synoptic influence. The three simulations differed in the strength of the background wind speed (but not in initial soil moisture). To analyze SM–P coupling, they considered the pixel-wise linear correlation between morning soil moisture and various atmospheric variables, including afternoon precipitation. They observed that convection is preferentially initiated over

drier patches, but that convective cells strengthen and preferentially precipitate when propagated over wet patches. They concluded that there is a weak negative SM–P coupling when the wind is too weak to propagate convective cells to wet patches, while there is a stronger positive SM–P coupling otherwise.

Welty and Zeng (2018) considered observational and reanalysis data of ten summer seasons over the US Southern Great Plains and studied how soil moisture affects the development of afternoon precipitation events after their initiation. To that purpose, they considered only days with afternoon precipitation (and no morning precipitation), partitioned these days into three dynamic regimes based on daily water vapor convergence, and computed for each resulting group of days the correlation between morning soil moisture and afternoon precipitation magnitude. They found that the sign of the correlation depends on the dynamic regime (positive for high dynamic regime, negative for low dynamic regime) and becomes insignificant when all regime days are considered together.

Holgate et al. (2019) considered observational data over Australia and computed pixel-wise linear correlations at different spatial scales between daily average soil moisture and next-day rainfall. They noted that the locality assumption in many statistical approaches for studying SM–P coupling (where soil moisture at a pixel is compared to precipitation at the same pixel) is problematic: for example the linear correlation between soil moisture and precipitation at a pixel is meaningless if the wind speed is too large with respect to the considered spatial and temporal scales of the analysis. Consequently, for their analysis, they filtered out all days for which this assumption was not valid because the wind speed was too large. They also filtered out all precipitation events with previous day precipitation exceeding 1 mm to account for the confounding effect of precipitation persistence and studied each season separately to account for the confounding effect of seasonality. Depending on the location, they found mainly positive or no correlations, apart from Austral winter when they also found slightly negative correlations.

Another prominent statistical study on SM–P coupling is (Guillod et al., 2015). They used observational data and various metrics in an attempt to reconcile previous findings of different signs of SM–P coupling. First, Guillod et al. (2015) considered the methodology from (Taylor et al., 2012) described above. Second, they compared the strength of early morning soil moisture anomalies on rain event days to early morning soil moisture anomalies on non-event days. Lastly, they compared the average standard deviation of early morning soil moisture anomalies in the surrounding area of an afternoon rain event to the average standard deviation of early morning soil moisture anomalies for non-rain events. They concluded that afternoon rain is more likely to occur during wet and heterogeneous soil moisture conditions, while being located over comparatively drier patches.

A similar study but with other observational data sets is (Hsu et al., 2017). They used a different metric than (Guillod et al., 2015) to characterize whether the location of a rain event was relatively dry or wet compared to its surroundings before the event. They found that afternoon rain is more likely to occur at patches that are relatively dry compared to their surroundings, but that this preference is weakened under wetter soil conditions and even reversed in extremely wet times. Moreover, they observed that the preference of afternoon rain to occur at patches that are relatively dry compared to their surroundings strengthens the stronger the soil moisture inhomogeneities are.

Ford et al. (2015) used in situ observational data from eleven summer seasons in the US Southern Great Plains. Considering only afternoon precipitation events with no precipitation in the morning and without synoptic forcing (as assessed by manual inspection), and binning early morning soil moisture values, they found a clear preference for afternoon precipitation events to occur over drier than median soils (which could indicate a negative SM–P coupling, but could also be related to the above-described preference of afternoon rain to occur at patches that are relatively dry compared to their surroundings). Ford et al. (2018) considered different remote sensing observational data over the US Great Plains and classified afternoon precipitation events into weakly or synoptically forced. Comparing morning soil moisture anomalies for days with and without afternoon precipitation events for the different data sets and groups of days, they identified different signs of SM–P coupling for different soil moisture data sets and for different convective environments (i.e. weakly or synoptically forced).

Graf et al. (2021) considered the output of convection-permitting simulations of the summer season 2016 for two regions, a prealpine, humid region in Southern Germany and a semiarid region in West Africa. Aligning the average magnitude of the soil moisture gradient in some area with next-hour precipitation in that area, they found a preference for precipitation to initiate over areas with high soil moisture gradients for the region in Southern Germany. Besides, they detected a preference for precipitation to initiate over dry areas in the region in Southern Germany, and over wet areas in the region in West Africa.

A particularly promising statistical approach for determining the actual *causal* impact (see Chapter 3) of soil moisture increases on precipitation was described in (Li et al., 2020; Tuttle and Salvucci, 2016, 2017). It uses Granger causality (see Chapter 3) to investigate the relation between soil moisture and the occurrence of next-day precipitation. The concept is to train two separate models with several input variables representing a set of processes that could influence subsequent precipitation occurrence. For one of the two models (the *full model*), soil moisture is included as an input variable while for the other (the *restricted model*) soil moisture is not included. Then, the scientist evaluates if the prediction capability of the full model is (significantly) better than that of the restricted model. For all locations where this is the case, the data set is divided into

days with higher than seasonal median soil moisture (*wet days*) and days with lower than seasonal median soil moisture (*dry days*). Next, the ratio of the precipitation probability predicted by the full model and by the restricted model, averaged over all wet and dry days, respectively, is computed. If this ratio is larger than one for wet days and smaller than one for dry days, the impact of soil moisture on precipitation is called positive, while it is called negative if the ratio is smaller than one for wet days and larger than one for dry days.

Applying this approach to observational data over the contiguous United States, Tuttle and Salvucci (2016) found that an increase in soil moisture at some pixel caused an increase in next-day precipitation probability at that pixel in arid regions (positive soil-moisture–precipitation-*probability* coupling), while it caused a decrease in next-day precipitation probability in more humid, vegetated areas. However, they remarked that the detected negative coupling might be erroneous due to uncertainties in the data in the respective region. Indeed, using various additional observational and reanalysis data sets, as well as non-linear instead of linear models, Li et al. (2020) confirmed the positive soil-moisture–precipitation-probability coupling over dry and transition zones, but not the negative coupling. Li et al. (2020) detected a particularly strong positive impact of soil moisture on next-day precipitation probability at the leeward slope of the Rocky Mountains, which they explained by water vapor being blocked by the mountains leading to evaporation being strongly controlled by soil moisture rather than by the horizontal transport of water vapor.

In summary, studies based on statistical approaches agree with the above findings from modelling studies that an increase in soil moisture tends to be associated with an increase in precipitation. However, most of them also indicate that an increase in soil moisture increases the probability of precipitation events, which seems to be in contrast to the results from the above modelling studies. Nevertheless, a strong dependence of the coupling signs on the considered data set, the synoptic situation, and the considered region have been reported. Studies based on statistical approaches agree with modelling studies that convection seems to initiate more often over soils that are relatively dry compared to their surrounding area. Altogether, current results on SM–P coupling are inconclusive and many aspects of the coupling remain poorly understood.

5. Variant approach for identifying spurious relations that deep learning models learn

This chapter summarizes the research article

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

The article is attached as Appendix A.

As one of the main research topics of this thesis, I evolve the recently proposed approach of using interpretable deep learning (DL) to gain new scientific insights into the Earth system. In this approach, a DL model is trained to predict one (target) variable given some other (input) variables. During training, the model learns a function relating the input and target variables. After training, an interpretation method is used to obtain descriptions of the learned function (e.g. representing the importance of different input variables for the predictions of the model; see Figure 1.2). Finally, these descriptions are analyzed to obtain new scientific insights. The major challenge to this approach is that it merely provides insights into statistical associations rather than actual causal relations between Earth system variables (see Chapter 3). Deciding whether relations that a DL model learned reflect mere statistical associations (referred to as *spurious correlations*) or actual causal relations is challenging because the relations between Earth system variables are often unknown or complex. In this article, I address this challenge developing a novel methodology to identify spurious correlations that a DL model learned.

In the proposed methodology (referred to as variant approach), separate instances of the considered DL model (referred to as variant models) are trained 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. Next, the descriptions of the functions that original and variant models learn are compared and it is evaluated whether they reflect the assumed stability or specific variation, respectively, of causal relations. If this is not the case for some parts of the descriptions, these parts likely reflect spurious correlations. The approach constitutes a generalization of sampling approaches, where separate instances of the considered DL model are trained on random samples of the training set and the obtained descriptions are compared or aggregated (Bin et al., 2015; Gagne II et al., 2019).

To illustrate the methodology, I consider two prediction tasks from hydrometeorology. In the first task, the occurrence of rain at a target location is predicted given geopotential fields at different pressure levels in a surrounding region. In the second task, the water level at a location in a river is predicted given the water level upstream and downstream 48 h earlier. As the proposed variant approach is not restricted to DL models, but valid for any statistical method, I demonstrate the approach using linear models (linear and logistic regression) as well as neural networks (multilayer perceptrons and convolutional neural networks (CNNs)). Note that, in the article, I also provide an introduction to all considered statistical models, in particular to CNNs, which are also used in Chapters 6 and 7 of this thesis. After training one of the statistical models on one of the prediction tasks, an interpretation method is applied to obtain a measure of the average importance of different input locations for the predictions of the model. Then, the variant approach is used to identify if this importance reflects spurious instead of causal relations between input and target

variables. I consider simple prediction tasks and simple descriptions of the learned functions to be able to decide whether parts of the descriptions that the variant approach identifies as spurious do indeed reflect spurious correlations. This is necessary to evaluate the variant approach. In the next chapters, I also apply the variant approach in my study of soil-moisture–precipitation (SM–P) coupling, which constitutes a highly complex relation.

The considered examples demonstrate the superiority of the proposed variant approach over pure sampling approaches. Indeed, the variant approach allows to identify spurious correlations that go unnoticed when applying pure sampling approaches. For the rain prediction task, where I assume causal relations to remain stable between original and variant tasks, the variant approach allows to correctly identify spurious correlations that the statistical models learned by formally evaluating the distances between original and variant descriptions. This means that the approach requires only minimal human intervention after the variant tasks are defined (e.g. adjusting a threshold).

In the water level prediction task, where formally specifying the assumed variation of causal relations is more involved, the variant approach also allows to identify spurious correlations that the statistical models learned. However, this task also demonstrates a challenge to the approach: when the assumed variation of causal relations in the variant tasks cannot be formalized, then the evaluation of distances between original and variant descriptions cannot be formalized either and requires expert visual assessment of original and variant descriptions.

An important aspect to consider when applying the variant approach is that it cannot guarantee that all spurious correlations reflected in a description are identified. For example, when considering variant tasks for which causal relations are assumed to remain stable, spurious correlations can only be identified if they do *not* remain stable between original and variant tasks. Thus, the variant approach can increase the confidence that a description reflects causal relations (or reject that this is the case), but it cannot guarantee it.

I developed the described methodology with contributions from Stefan Kollet and Jochen Garcke. The experiments described in the article were designed by me with suggestions by Stefan Kollet. I conducted all experiments. I analyzed the results and prepared the manuscript with contributions from Stefan Kollet and Jochen Garcke. All co-authors agreed to the use of the article in this thesis and the description of author contributions.

6. Causal deep learning models for studying the Earth system

This chapter summarizes the research article

T. Tesch, S. Kollet, and J. Garcke. Causal deep learning models for studying the Earth system. *Geoscientific Model Development*, 16(8):2149–2166, 2023a. doi: 10.5194/gmd-16-2149-2023. Highlight article.

The article is attached as Appendix B.

The major challenge to the approach of using interpretable deep learning to gain new scientific insights is that DL models learn statistical associations rather than causal relations between Earth system variables. This article considers modifications of the prediction task and training procedure to obtain a *causal* DL model, i.e. a model that actually learns the *causal* relation between some input variable and the target variable. Specifically, in this article, I combine the approach of using interpretable DL to gain new scientific insights with a result from causality research (Pearl, 2009) stating that a statistical model may learn the causal impact of an input variable on a target variable if suitable additional input variables are chosen to prevent confounding (see also Chapter 3). In contrast to the variant approach described in the previous chapter, which considers just any already trained DL model and addresses the challenge of identifying spurious correlations that the model learned, the methodology described in this chapter focuses on obtaining a causal DL model by modifying the prediction task and training procedure (see Figure 1.2).

The article begins with an introduction to the framework of structural causal models (SCMs) (Pearl, 2009), which provides notation and concepts for formalizing research on causal relations. In particular, it introduces causal graphs and structural equations, which are the basis for describing a (sub)system of interest (e.g. part of the Earth system) within the framework of SCMs. Furthermore, it introduces the do-operator, which represents arbitrary interventions, here into the Earth system, and is used to formalize the notion of causal effects. In this framework, variables have a deterministic and a random component (see Figure 3.2). The randomness represents random aspects of the system like turbulence and aspects of the system that are not modelled explicitly, for example due to the considered spatial scale.

After the introduction to SCMs, a causal DL model is defined as a DL model that approximates the map

$$(x, \{c_\ell\}_{\ell=1}^k) \to \mathbb{E}[Y|do(X = x), \{C_\ell = c_\ell\}_{\ell=1}^k], \tag{6.1}$$

where $Y \in \mathbb{R}^n$ is the considered target variable, $X \in \mathbb{R}^d$ is the input variable of interest, and $\{C_\ell\}_{\ell=1}^k$ are additional input variables. Note that small letters x, y and c_ℓ refer to particular values of the random variables X, Y and C_ℓ, respectively.

The expression $do(X = x)$ distinguishes a causal DL model from a standard DL model. It represents an arbitrary intervention into the considered system, such that the term on the right hand side of Equation 6.1 is the expected value of Y given the variables $\{C_\ell\}_{\ell=1}^k$ and given that one intervened into the system and set X to some arbitrary value x (as one could do in a real experiment or when using numerical models of the considered system). Obtaining a causal DL model requires a careful choice of loss function, DL model and additional input variables, which is described in detail in this article.

Given a causal DL model, the partial derivatives of the model approximate the partial derivatives of the map from Equation 6.1, i.e.

$$s_{ij}(x, \{c_\ell\}_{\ell=1}^k) = \frac{\partial \mathbb{E}[Y_i | do(X = x), \{C_\ell = c_\ell\}_{\ell=1}^k]}{\partial X_j}, \tag{6.2}$$

where $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, d\}$. These partial derivatives describe how Y_i changes if one intervened into the system and slightly changed the value of X_j, i.e. they represent the causal impact of X_j on Y_i. They may be averaged over several input samples $(x, \{c_\ell\}_{\ell=1}^k)$ to obtain the average impact of a change in X_j on Y_i, and aggregated over several indexes i and j to obtain for example the impact of a change in X_j on the sum $\sum_{i=1}^n Y_i$.

In addition to the methodology itself, I propose several further analyses to assess whether results obtained with the methodology are statistically significant, i.e. reflect more than random correlations or artifacts of the DL training procedure, and whether they reflect more than known correlations. Further, I propose the variant approach from Chapter 5 to assess whether the obtained results reflect (potentially unknown) spurious correlations rather than actual causal relations.

The methodology itself as well as the additional analyses are illustrated with the example of soil-moisture–precipitation (SM–P) coupling in ERA5 data across Europe. The impact of soil moisture on subsequent precipitation is highly complex and remains poorly understood despite decades of research (see Chapter 4). My results indicate that a local increase in soil moisture leads to a local increase in precipitation, but to a regional decrease in precipitation. This effect seems to be enhanced by mountainous regions and ridges. As the focus of this article is the methodology, further investigation and discussion of these results follow in another article (see Chapter 7).

The obtained results on SM–P coupling differ entirely from results obtained with a simple linear correlation analysis between soil moisture and precipitation. This stresses the importance of taking into account the difference between correlation and causation and the importance of using statistical models that can represent the non-linearity of relations in the Earth system.

I developed the described methodology with contributions from Stefan Kollet and Jochen Garcke. I conducted the experiments for the illustrative example of SM–P coupling. I analyzed the results and prepared the manuscript with contributions from Stefan Kollet and Jochen Garcke. All co-authors agreed to the use of the article in this thesis and the description of author contributions.

7. Converse local and non-local soil-moisture–precipitation couplings across Europe

This chapter summarizes a research article ready for submission to a scientific journal. The authors of the article are Tobias Tesch, Stefan Kollet, Jochen Garcke, Stergios Kartsios, and Eleni Katragkou. The article is attached as Appendix C.

In this article, I apply the methodology of causal deep learning (DL) models, developed in the previous chapter, to study the impact of soil moisture changes on subsequent precipitation across Europe at a sub-daily time scale. Two different data sets are used for this analysis: first, ERA5 climate reanalysis data (Hersbach et al., 2018) across Europe, which are a reanalysis of the past decades (1950 to today) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF) and contain hourly estimates of a large number of Earth system variables on a regular latitude-longitude grid of 0.25 degrees (≈ 30 km). Reanalysis means that they combine simulation data and observations into a single description of the global climate and weather to obtain estimates of the considered Earth system variables that are as close to reality as possible. Second, I consider data from a high-resolution, convection-permitting simulation covering the years 2000 to 2014 and containing hourly estimates of several Earth system variables on a rotated latitude-longitude grid of 0.0275 degrees (≈ 3 km) across central Europe (Tesch et al., 2022b), hereafter referred to by CP data. While this is no reanalysis and the estimates of the considered Earth system variables might be less close to reality than in the ERA5 data, the high-resolution, convection-permitting simulation represents processes more realistically that are essential for soil-moisture–precipitation (SM–P) coupling, e.g. convection and thermally driven circulations (Leutwyler et al., 2021).

In terms of methodology, in this article, I describe a more efficient way of computing the required partial derivatives of the DL models, because the naive way for computing these derivatives is computationally infeasible for the CP data. Further, I add an analysis related to Granger causality (see Chapter 3), where the performance of the original DL model is compared to the performance of a DL model trained on permuted soil moisture data to assess whether the original DL model learned useful information on SM–P coupling in terms of predictive performance apart from noise, and the correlations between SM and topography or seasonality (which are preserved by the considered soil moisture permutations).

The obtained results confirm the results from the illustrative example in the previous chapter: a local increase in soil moisture leads to a local increase in precipitation, but to a regional decrease in precipitation, and this effect is enhanced by mountainous regions and ridges. Note that the latter is *not* due to enhanced precipitation in mountainous regions, but actually due to a stronger impact of soil moisture changes on precipitation. Furthermore, the average impact of local soil moisture changes on local and regional precipitation is found to be qualitatively very similar to the average impact on local and regional precipitation *probability*. Although the results for ERA5 data and CP data are qualitatively similar, the obtained SM–P couplings for the CP data are not significant at many pixels in the considered region, i.e. do not differ significantly from the couplings obtained for models trained on permuted soil moisture data. I believe that this is due to the highly chaotic nature of convection (Leutwyler et al., 2021), and because less training years were available for

the CP data than for the ERA5 data, while the considered DL model even had more parameters. These aspects make it more difficult for the DL model to learn correct SM–P coupling for the CP data than for the ERA5 data.

In the article, the results are discussed in light of related studies (see also Chapter 4). The obtained positive local SM–P coupling is in line with most previous statistical and modelling studies on SM–P coupling, which found that precipitation tends to increase in regions where soil moisture is increased. Concerning the obtained negative non-local SM–P coupling, there is only one study (Imamovic et al., 2017) considering this question on a similar spatial scale. They found that local precipitation was reduced for a local increase in soil moisture (negative local coupling), while non-local precipitation was less affected. This is in contrast to my findings on a positive local and a negative non-local SM–P coupling, while it agrees with an overall (regional) negative SM–P coupling. Note however, that Imamovic et al. (2017) considered a different time scale, and specific atmospheric, topographic, and initial soil moisture conditions. My findings on enhanced SM–P coupling in mountainous regions agree with findings in (Leutwyler et al., 2021) who found particularly strong SM–P coupling in the Alpine region, and (Li et al., 2020) who found particularly strong SM–P coupling at the leeward slope of the Rocky Mountains, while it is again in contrast to findings in (Imamovic et al., 2017). An important result from this study for future research on SM–P coupling is the importance of non-local effects in the coupling, which cause converse signs in local and regional SM–P couplings in my experiments and have commonly been neglected in previous studies.

I developed the faster implementation of the methodology of causal DL models and the additional significance analysis. I conducted all experiments. I analyzed the results and prepared the manuscript with contributions from Stefan Kollet and Jochen Garcke. Stergios Kartsios and Eleni Katragkou performed the convection-permitting simulations. All co-authors agreed to the use of the article in this thesis and the description of author contributions.

8. Conclusion and outlook

8.1. Summary

In this thesis, I evolved the recently proposed approach of using interpretable deep learning (DL) to gain new scientific insights into the Earth system. In particular, I addressed the main challenge to the approach, which is that DL models may learn spurious correlations rather than causal relations between input and target variables.

In a first step, I proposed a variant approach to identify spurious correlations that any given statistical model learned, and illustrated the superiority of the approach over commonly applied sampling approaches using two examples from hydrometeorology in combination with various statistical models. The variant approach allowed to identify various spurious correlations that the models learned, while some of these relations went unnoticed in commonly applied sampling approaches.

Next, I developed the methodology of causal DL models. It combines the approach of using interpretable DL to gain new scientific insights with findings from causality research in order to achieve that a DL model actually learns causal relations between input and target variables rather than spurious correlations. Moreover, I proposed several analyses to assess the correctness of results obtained with this methodology. These analyses include, but are not limited to, the aforementioned variant approach. I applied the developed methodology to study soil-moisture–precipitation (SM–P) coupling obtaining results that differ substantially from results obtained with a simple linear correlation analysis. This underlines that the considerations made in the methodology, in particular on the difference between correlation and causation, and the use of statistical models that can represent non-linear relations, are crucial to obtain new scientific insights on SM–P coupling. These considerations and the proposed methodology are also useful to obtain new insights into other complex relations in the Earth system.

After developing and illustrating the methodology of causal DL models, I applied it to gain new scientific insights on SM–P coupling. Applying the methodology to reanalysis data and data from a high-resolution, convection-permitting simulation, I found a local increase in soil moisture to cause an increase in subsequent local precipitation, but an even stronger decrease in subsequent non-local precipitation. Both effects are enhanced by mountainous regions and ridges. Further, I found the average impact of local soil moisture changes on local and regional precipitation (amount) to be qualitatively very similar to the average impact on local and regional precipitation probability. In particular, these results stress the importance of taking into account non-local effects, which have mostly been neglected in previous works, in future studies on SM–P coupling.

8.2. Challenges and limitations

In terms of methodology, the fact that DL models may learn spurious correlations rather than causal relations between input and target variables remains challenging despite the progress made in this thesis. While the proposed variant approach has proven useful in identifying spurious correlations that a given DL model learned, it cannot guarantee that the model learned causal relations, such that some uncertainty concerning the correctness of the learned relations remains. Quantifying this uncertainty seems extremely challenging as well.

The developed methodology of causal DL models, in theory, ensures that a DL model learns causal rather than spurious correlations between input and target variables. However, in practice, this is only the case if no confounding variables are neglected in the choice of input variables, i.e. if the input variables fulfil the adjustment criteria from (Shpitser et al., 2010), and if the DL model provides a perfect approximation of the expected value of the considered target variable given all input variables (i.e. of the function in Equation 6.1). Neither is realistic: the former due to the complexity of the Earth system with its large number of spatially and temporally continuous variables, and the latter due to common problems of DL, such as overfitting, i.e. making good predictions on the training set but for wrong reasons. Furthermore, the former aspect favors including a large number of input variables, while the latter penalizes the number of input variables (see Figure 8.1). Indeed, additional input variables increase the complexity of the approximated expected value, and increase the general risk of overfitting. The choice of input variables is further complicated because both aspects, i.e. the error in the learned relations due to neglected confounders as well as the error due to the approximation of the considered expected value, are challenging to quantify.

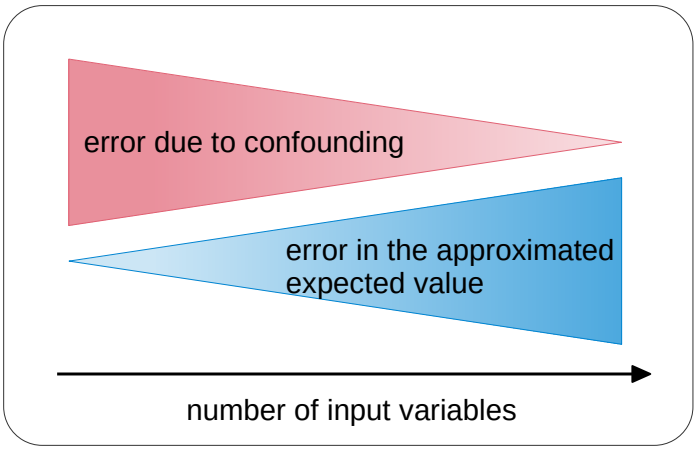

Figure 8.1: Schematic of the errors in the methodology of causal deep learning models. When more input variables are considered, the error due to confounding decreases, but the error in the approximated expected value increases.

In this thesis, I performed several further analyses increasing the confidence in my results, e.g. finding with high confidence that the results do not only reflect spurious correlations between soil moisture and topography or seasonality. However, some uncertainty remains, and again, quantifying this uncertainty is extremely challenging. It is important to note that these challenges apply not only to the proposed methodology, but that any other approach for studying the highly complex nature of the Earth system brings its own challenges and uncertainties. For example, approaches based on numerical simulations rely on temporal and spatial discretizations and a large number of parameterized processes based on simplifying assumptions (e.g. vegetation processes or micro-physical processes in clouds). The resulting uncertainties are very challenging to quantify as well. Meanwhile, other statistical approaches face the same challenges concerning for example the difference between correlation and causation (although the existence of this challenge is often ignored). Avoiding these model and statistical uncertainties by using interventional in-situ experiments in the Earth system is in many cases infeasible.

In terms of SM–P coupling, an additional difficulty is the chaotic nature of convection and the low signal-to-noise-ratio. Both aspects complicate learning SM–P coupling for a DL model. Further, they complicate assessing whether a DL model learned more than spurious correlations. Indeed, to that purpose, one of the considered analyses compared the performance of the original DL model to the performance of a separate instance of the model trained on permuted soil moisture data (hereafter referred to as variant model). While the performance averaged over the entire target region was significantly better for the original model than for the variant model, the absolute performance improvement was small. Furthermore, comparing the performance of the original model and the variant model for each target pixel separately revealed chaotic patterns showing an increase in performance for some pixels, but also a decrease for other pixels. The fraction of pixels showing an increase in performance for the original model increased with an increasing number of considered test years. This indicates that the original model is equally good or better than the variant model for all pixels, but that this is masked by the chaotic nature of convection and the low signal-to-noise-ratio. More test years would have been required to uncover these improvements.

8.3. Recommendations for future work

Regarding the presented results on SM–P coupling, some uncertainty remains due to the considered data sets and methodology (see previous section). Future works should further corroborate or reject these results. An important finding from this thesis for future research on SM–P coupling is that non-local effects are essential for the overall coupling and should not be neglected as it was the case in most previous studies.

In terms of methodology, there are several avenues for future work: first, additional methodologies

beyond the variant approach and the analyses described in Chapters 6 and 7 should be developed to identify spurious correlations that a DL model learned. Moreover, methods should be explored to quantify the uncertainty with that a learned relation is classified as causal or spurious. Some recently proposed methods for analyzing the uncertainty in *predictions* of DL models (rather than in the relations that these models learned) (Blundell et al., 2015; Lakshminarayanan et al., 2017; Loquercio et al., 2020) might be adapted for this purpose.

Another avenue for future research with respect to the developed methodology of causal DL models is the adaptation of other approaches for estimating causal effects. While regression adjustment is considered in this thesis, there are also more sophisticated approaches, e.g. based on propensity scores. An adaptation of these approaches to the cases considered here is not straightforward due to the continuous and high dimensional variables (as opposed to, for example, binary or discrete variables when studying the effects of medical treatments), as well as the complexity of the Earth system, but it might still be worthwhile.

Considering applications of the proposed methodologies, there are numerous opportunities for future studies. For instance, it would be interesting to study SM–P coupling in different data sets, different regions, or at different time scales. Apart from this, there are countless other complex relations in the Earth system that may be studied with the proposed methodology. To name a few examples, there are soil-moisture–temperature coupling (Schumacher et al., 2019; Schwingshackl et al., 2017; Seneviratne et al., 2006), soil-moisture–atmospheric-carbon-dioxide-coupling (Green et al., 2019; Humphrey et al., 2021), evaporation–precipitation coupling (Findell et al., 2011), snow-cover–precipitation coupling (Wallace and Minder, 2021), vegetation-cover–convective-boundary-layer coupling (Fisch et al., 2004), and groundwater–atmospheric-boundary-layer, groundwater–convective-available-potential-energy, and groundwater–precipitation coupling (Rahman et al., 2015).

In this work, I applied the developed methodologies to SM–P coupling in order to gain new scientific insights. Another application could be for validating if the coupling is correctly represented in a given numerical model. To that purpose, one could investigate differences in SM–P coupling between different numerical models, or between numerical models and observational data. Li et al. (2020), for example, proceeded similarly to study the ability of different numerical models to identify regions on Earth with particularly strong SM–P coupling. A conceptually similar approach for such process-oriented model validation is given in (Nowack et al., 2020). They apply a causal discovery algorithm to obtain a causal graph of atmospheric interactions in model simulations. To validate a model, they compare the causal graph obtained for that model to the causal graph obtained for reanalysis data, as a proxy for observations.

Concerning applications of the proposed variant approach, it would be interesting to further evaluate its potential for automatically identifying spurious correlations that DL models learn. Lapuschkin et al. (2019), for example, found that their DL model predicted that a given image shows a horse if it contains a certain copyright tag. This was because many horse images in the considered training set contained this copyright tag. The variant approach could help to uncover such flawed behavior of DL models. In the example from (Lapuschkin et al., 2019), a variant task could be to train separate instances of the considered DL model (*variant models*) on subsets of the data coming from different sources. If the copyright tag was not present in one of these sources, the faulty behavior could be uncovered by automatically comparing descriptions of the functions (e.g. feature importance scores) that the different variant models learn. In general, the proposed variant approach may be useful for debugging a DL model whenever the user cannot judge whether an explanation obtained by an interpretation method reflects undesired behavior of the DL model or not, i.e. when the relations between input and target variables are complex or unknown. This is a common case across scientific disciplines.

Concluding, this thesis demonstrated the potential of combining DL and causality research to gain new scientific insights into the Earth system. The developed methodologies provided important insights into soil-moisture–precipitation coupling and have several further promising applications in the geosciences and beyond. This thesis calls for more research on the combination of DL and causality research, as well as on applications of causal inference methods in the geosciences.

References

B. Adler, N. Kalthoff, and L. Gantner. Initiation of deep convection caused by land-surface inhomogeneities in West Africa: a modelled case study. *Meteorology and Atmospheric Physics*, 112 (1-2):15–27, 2011. doi: 10.1007/s00703-011-0131-2.

S. Agrawal, L. Barrington, C. Bromberg, J. Burge, C. Gazen, and J. Hickey. Machine Learning for Precipitation Nowcasting from Radar Images. *arXiv*, 2019. doi: 10.48550/arxiv.1912.12132.

F. Aires, P. Gentine, K. L. Findell, B. R. Lintner, and C. Kerr. Neural Network–Based Sensitivity Analysis of Summertime Convection over the Continental United States. *Journal of Climate*, 27 (5):1958–1979, 2014. doi: 10.1175/jcli-d-13-00161.1.

A. Amidi and S. Amidi. Convolutional Neural Networks cheatsheet. `https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks`. Accessed: 2023-05-19.

P. Autier and S. Gandini. Vitamin D Supplementation and Total Mortality: A Meta-analysis of Randomized Controlled Trials. *Archives of Internal Medicine*, 167(16):1730–1737, 2007. doi: 10.1001/archinte.167.16.1730.

S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, 2015. doi: 10.1371/journal.pone.0130140.

V. Badrinarayanan, A. Kendall, and R. Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12):2481–2495, 2017. doi: 10.1109/TPAMI.2016.2644615.

N. Ban, C. Caillaud, E. Coppola, E. Pichelli, S. Sobolowski, M. Adinolfi, B. Ahrens, A. Alias, I. Anders, S. Bastin, D. Belušić, S. Berthou, E. Brisson, R. M. Cardoso, S. C. Chan, O. B. Christensen, J. Fernández, L. Fita, T. Frisius, G. Gašparac, F. Giorgi, K. Goergen, J. E. Haugen, Ø. Hodnebrog, S. Kartsios, E. Katragkou, E. J. Kendon, K. Keuler, A. Lavin-Gullon, G. Lenderink, D. Leutwyler, T. Lorenz, D. Maraun, P. Mercogliano, J. Milovac, H.-J. Panitz, M. Raffa, A. R. Remedio, C. Schär, P. M. M. Soares, L. Srnec, B. M. Steensen, P. Stocchi, M. H. Tölle, H. Truhetz, J. Vergara-Temprado, H. de Vries, K. Warrach-Sagi, V. Wulfmeyer, and M. J. Zander. The first multi-model ensemble of regional climate simulations at kilometer-scale resolution, part I: evaluation of precipitation. *Climate Dynamics*, 57(1-2):275–302, 2021. doi: 10.1007/s00382-021-05708-w.

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

C. Barthlott and N. Kalthoff. A Numerical Sensitivity Study on the Impact of Soil Moisture on Convection-Related Parameters and Convective Precipitation over Complex Terrain. *Journal of the Atmospheric Sciences*, 68(12):2971–2987, 2011. doi: 10.1175/jas-d-11-027.1.

F. Baur, C. Keil, and G. C. Craig. 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(717):2863–2875, 2018. doi: 10.1002/qj.3415.

R. D. Bin, S. Janitza, W. Sauerbrei, and A.-L. Boulesteix. Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics*, 72(1):272–280, 2015. doi: 10.1111/biom.12381.

C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight Uncertainty in Neural Networks. In *32nd International Conference on Machine Learning, PMLR*, volume 37, pages 1613–1622, 2015. doi: 10.48550/arXiv.1505.05424.

W. Brutsaert. *Hydrology: An Introduction*. Cambridge University Press, 2005. doi: 10.1017/CBO9780511808470.

G. Camps-Valls, M. Reichstein, X. Zhu, and D. Tuia. Advancing deep learning for Earth sciences: from hybrid modeling to interpretability. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 3979–3982, 2020. doi: 10.1109/IGARSS39084.2020.9323558.

G. Camps-Valls, D. Tuia, X. X. Zhu, and M. Reichstein, editors. *Deep Learning for the Earth Sciences*. Wiley, 2021. doi: 10.1002/9781119646181.

M. S. Cepeda. Comparison of Logistic Regression versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders. *American Journal of Epidemiology*, 158(3): 280–287, 2003. doi: 10.1093/aje/kwg115.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

G. Cioni and C. Hohenegger. Effect of Soil Moisture on Diurnal Convection and Precipitation in Large-Eddy Simulations. *Journal of Hydrometeorology*, 18(7):1885–1903, 2017. doi: 10.1175/jhm-d-16-0241.1.

E. Coppola, S. Sobolowski, E. Pichelli, F. Raffaele, B. Ahrens, I. Anders, N. Ban, S. Bastin, M. Belda, D. Belusic, A. Caldas-Alvarez, R. M. Cardoso, S. Davolio, A. Dobler, J. Fernandez, L. Fita, Q. Fumiere, F. Giorgi, K. Goergen, I. Güttler, T. Halenka, D. Heinzeller, Ø. Hodnebrog, D. Jacob, S. Kartsios, E. Katragkou, E. Kendon, S. Khodayar, H. Kunstmann, S. Knist, A. Lavín-Gullón, P. Lind, T. Lorenz, D. Maraun, L. Marelle, E. van Meijgaard, J. Milovac, G. Myhre, H.-J. Panitz, M. Piazza, M. Raffa, T. Raub, B. Rockel, C. Schär, K. Sieck, P. M. M. Soares, S. Somot, L. Srnec, P. Stocchi, M. H. Tölle, H. Truhetz, R. Vautard, H. de Vries, and K. Warrach-Sagi. A first-of-its-kind multi-model convection permitting ensemble for investigating convective phenomena over Europe and the Mediterranean. *Climate Dynamics*, 55(1-2):3–34, 2018. doi: 10.1007/s00382-018-4521-8.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989. doi: 10.1007/bf02551274.

A. Daw, A. Karpatne, W. Watkins, J. Read, and V. Kumar. Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling. *arXiv*, 2017. doi: 10.48550/arxiv.1710.11431.

L. Deng and D. Yu. Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014. doi: 10.1561/2000000039.

J. Dramsch, G. Corte, H. Amini, C. MacBeth, and M. Lüthje. Including Physics in Deep Learning – An Example from 4D Seismic Pressure Saturation Inversion. In *81st EAGE Conference and Exhibition 2019 Workshop Programme*. European Association of Geoscientists & Engineers, 2019. doi: 10.3997/2214-4609.201901967.

V. Dumoulin and F. Visin. A guide to convolution arithmetic for deep learning. *arXiv*, 2016. doi: 10.48550/arXiv.1603.07285.

I. Ebert-Uphoff and Y. Deng. Causal Discovery for Climate Research Using Graphical Models. *Journal of Climate*, 25(17):5648–5665, 2012. doi: 10.1175/jcli-d-11-00387.1.

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

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

E. A. B. Eltahir. A Soil Moisture–Rainfall Feedback Mechanism: 1. Theory and observations. *Water Resources Research*, 34(4):765–776, 1998. doi: 10.1029/97WR03499.

M. C. Elze, J. Gregson, U. Baber, E. Williamson, S. Sartori, R. Mehran, M. Nichols, G. W. Stone, and S. J. Pocock. Comparison of Propensity Score Methods and Covariate Adjustment. *Journal of the American College of Cardiology*, 69(3):345–357, 2017. doi: 10.1016/j.jacc.2016.10.060.

L. Espeholt, S. Agrawal, C. Sønderby, M. Kumar, J. Heek, C. Bromberg, C. Gazen, R. Carver, M. Andrychowicz, J. Hickey, A. Bell, and N. Kalchbrenner. Deep learning for twelve hour precipitation forecasts. *Nature Communications*, 13(1), 2022. doi: 10.1038/s41467-022-32483-x.

K. L. Findell and E. A. B. Eltahir. Atmospheric Controls on Soil Moisture–Boundary Layer Interactions. Part I: Framework Development. *Journal of Hydrometeorology*, 4(3):552–569, 2003a. doi: 10.1175/1525-7541(2003)004<0552:acosml>2.0.co;2.

K. L. Findell and E. A. B. Eltahir. Atmospheric Controls on Soil Moisture–Boundary Layer Interactions. Part II: Feedbacks within the Continental United States. *Journal of Hydrometeorology*, 4 (3):570–583, 2003b. doi: 10.1175/1525-7541(2003)004<0570:acosml>2.0.co;2.

K. L. Findell, P. Gentine, B. R. Lintner, and C. Kerr. Probability of afternoon precipitation in eastern United States and Mexico enhanced by high evaporation. *Nature Geoscience*, 4(7):434–439, 2011. doi: 10.1038/ngeo1174.

G. Fisch, J. Tota, L. A. T. Machado, M. A. F. S. Dias, R. F. da F. Lyra, C. A. Nobre, A. J. Dolman, and J. H. C. Gash. The convective boundary layer over pasture and forest in Amazonia. *Theoretical and Applied Climatology*, 78(1-3), 2004. doi: 10.1007/s00704-004-0043-x.

T. W. Ford, A. D. Rapp, S. M. Quiring, and J. Blake. Soil moisture–precipitation coupling: observations from the Oklahoma Mesonet and underlying physical mechanisms. *Hydrology and Earth System Sciences*, 19(8):3617–3631, 2015. doi: 10.5194/hess-19-3617-2015.

T. W. Ford, S. M. Quiring, B. Thakur, R. Jogineedi, A. Houston, S. Yuan, A. Kalra, and N. Lock. Evaluating Soil Moisture–Precipitation Interactions Using Remote Sensing: A Sensitivity Analysis. *Journal of Hydrometeorology*, 19(8):1237–1253, 2018. doi: 10.1175/jhm-d-17-0243.1.

P. Froidevaux, L. Schlemmer, J. Schmidli, W. Langhans, and C. Schär. Influence of the Background Wind on the Local Soil Moisture–Precipitation Feedback. *Journal of the Atmospheric Sciences*, 71(2):782–799, 2014. doi: 10.1175/jas-d-13-0180.1.

C. Furusho-Percot, K. Goergen, C. Hartick, K. Kulkarni, J. Keune, and S. Kollet. Pan-European groundwater to atmosphere terrestrial systems climatology from a physically consistent simulation. *Scientific Data*, 6(1):320, 2019. doi: 10.1038/s41597-019-0328-7.

D. J. Gagne II, S. E. Haupt, D. W. Nychka, and G. Thompson. Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms. *Monthly Weather Review*, 147(8):2827–2845, 2019. doi: 10.1175/mwr-d-18-0316.1.

D. Galagate. Causal inference with a continuous treatment and outcome: alternative estimators for parametric dose-response functions with applications. *Digital Repository at the University of Maryland*, 2016. doi: 10.13016/M2Q48K. PhD thesis.

F. Gasper, K. Goergen, P. Shrestha, M. Sulis, J. Rihani, M. Geimer, and S. Kollet. Implementation and scaling of the fully coupled Terrestrial Systems Modeling Platform (TerrSysMP v1.0) in a massively parallel supercomputing environment – a case study on JUQUEEN (IBM Blue Gene/Q). *Geoscientific Model Development*, 7(5):2531–2543, 2014. doi: 10.5194/gmd-7-2531-2014.

P. Gentine, A. A. M. Holtslag, F. D'Andrea, and M. Ek. Surface and Atmospheric Controls on the Onset of Moist Convection over Land. *Journal of Hydrometeorology*, 14(5):1443–1462, 2013. doi: 10.1175/jhm-d-12-0137.1.

P. Gentine, A. Massmann, B. R. Lintner, S. H. Alemohammad, R. Fu, J. K. Green, D. Kennedy, and J. Vilà-Guerau de Arellano. Land–atmosphere interactions in the tropics – a review. *Hydrology and Earth System Sciences*, 23(10):4171–4197, 2019. doi: 10.5194/hess-23-4171-2019.

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

I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, 2015. doi: 10.48550/arXiv.1412.6572.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

M. Graf, J. Arnault, B. Fersch, and H. Kunstmann. Is the soil moisture precipitation feedback enhanced by heterogeneity and dry soils? A comparative study. *Hydrological Processes*, 35(9), 2021. doi: 10.1002/hyp.14332.

C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, 1969. doi: 10.2307/1912791.

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

J. K. Green, S. I. Seneviratne, A. M. Berg, K. L. Findell, S. Hagemann, D. M. Lawrence, and P. Gentine. Large influence of soil moisture on long-term terrestrial carbon uptake. *Nature*, 565 (7740):476–479, 2019. doi: 10.1038/s41586-018-0848-x.

B. P. Guillod, B. Orlowsky, D. Miralles, A. J. Teuling, P. D. Blanken, N. Buchmann, P. Ciais, M. Ek, K. L. Findell, P. Gentine, B. R. Lintner, R. L. Scott, B. V. den Hurk, and S. I. Seneviratne. Land-surface controls on afternoon precipitation diagnosed from observational data: uncertainties and confounding factors. *Atmospheric Chemistry and Physics*, 14(16):8343–8367, 2014. doi: 10.5194/acp-14-8343-2014.

B. P. Guillod, B. Orlowsky, D. G. Miralles, A. J. Teuling, and S. I. Seneviratne. Reconciling spatial and temporal soil moisture effects on afternoon rainfall. *Nature Communications*, 6(1), 2015. doi: 10.1038/ncomms7443.

R. Guo, L. Cheng, J. Li, P. R. Hahn, and H. Liu. A Survey of Learning Causality with Data. *ACM Computing Surveys*, 53(4):1–37, 2021. doi: 10.1145/3397269.

Y. Guo, X. Cao, B. Liu, and M. Gao. Cloud Detection for Satellite Imagery Using Attention-Based U-Net Convolutional Neural Network. *Symmetry*, 12(6):1056, 2020. doi: 10.3390/sym12061056.

Z. Guo, P. A. Dirmeyer, R. D. Koster, Y. C. Sud, G. Bonan, K. W. Oleson, E. Chan, D. Verseghy, P. Cox, C. T. Gordon, J. L. McGregor, S. Kanae, E. Kowalczyk, D. Lawrence, P. Liu, D. Mocko, C.-H. Lu, K. Mitchell, S. Malyshev, B. McAvaney, T. Oki, T. Yamada, A. Pitman, C. M. Taylor, R. Vasic, and Y. Xue. GLACE: The Global Land–Atmosphere Coupling Experiment. Part II: Analysis. *Journal of Hydrometeorology*, 7(4):611–625, 2006. doi: 10.1175/jhm511.1.

Y.-G. Ham, J.-H. Kim, and J.-J. Luo. Deep learning for multi-year ENSO forecasts. *Nature*, 573 (7775):568–572, 2019. doi: 10.1038/s41586-019-1559-7.

L. Han, H. Liang, H. Chen, W. Zhang, and Y. Ge. Convective Precipitation Nowcasting Using U-Net Model. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–8, 2022. doi: 10.1109/TGRS.2021.3100847.

E. Hariton and J. J. Locascio. Randomised controlled trials - the gold standard for effectiveness research. *BJOG: Int J Obstet Gy*, 125(13):1716–1716, 2018. doi: 10.1111/1471-0528.15199.

C. Hartick, C. Furusho-Percot, K. Goergen, and S. Kollet. An Interannual Probabilistic Assessment of Subsurface Water Storage Over Europe Using a Fully Coupled Terrestrial Model. *Water Resources Research*, 57(1):e2020WR027828, 2021. doi: 10.1029/2020WR027828.

C. Hauck, C. Barthlott, L. Krauss, and N. Kalthoff. Soil moisture variability and its influence on convective precipitation over complex terrain. *Quarterly Journal of the Royal Meteorological Society*, 137(S1):42–56, 2011. doi: 10.1002/qj.766.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

L. Henckel, E. Perković, and M. H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *Journal of the Royal Statistical Society Series B*, 84(2): 579–599, 2022. doi: 10.1111/rssb.12451.

O. Henneberg, F. Ament, and V. Grützun. Assessing the uncertainty of soil moisture impacts on convective precipitation using a new ensemble approach. *Atmospheric Chemistry and Physics*, 18(9):6413–6425, 2018. doi: 10.5194/acp-18-6413-2018.

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

T. Hesterberg. What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. *arXiv*, 2014. doi: 10.48550/arXiv.1411.5279.

J. L. Hill. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162.

C. Hohenegger, P. Brockhaus, C. S. Bretherton, and C. Schär. The Soil Moisture–Precipitation Feedback in Simulations with Explicit and Parameterized Convection. *Journal of Climate*, 22 (19):5003–5020, 2009. doi: 10.1175/2009jcli2604.1.

C. M. Holgate, A. I. J. M. V. Dijk, J. P. Evans, and A. J. Pitman. The Importance of the One-Dimensional Assumption in Soil Moisture - Rainfall Depth Correlation at Varying Spatial Scales. *Journal of Geophysical Research: Atmospheres*, 124(6):2964–2975, 2019. doi: 10.1029/2018jd029762.

K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991. doi: 10.1016/0893-6080(91)90009-t.

H. Hsu, M.-H. Lo, B. P. Guillod, D. G. Miralles, and S. Kumar. Relation between precipitation location and antecedent/subsequent soil moisture spatial patterns. *Journal of Geophysical Research: Atmospheres*, 122(12):6319–6328, 2017. doi: 10.1002/2016jd026042.

V. Humphrey, A. Berg, P. Ciais, P. Gentine, M. Jung, M. Reichstein, S. I. Seneviratne, and C. Frankenberg. Soil moisture–atmosphere feedback dominates land carbon uptake variability. *Nature*, 592(7852):65–69, 2021. doi: 10.1038/s41586-021-03325-5.

A. Imamovic, L. Schlemmer, and C. Schär. Collective Impacts of Orography and Soil Moisture on the Soil Moisture-Precipitation Feedback. *Geophysical Research Letters*, 44(22):11,682–11,691, 2017. doi: 10.1002/2017GL075657.

G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

S. Ioffe and C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *32nd International Conference on Machine Learning, PMLR*, volume 37, pages 448–456, 2015. doi: 10.48550/arXiv.1502.03167.

K. Kashinath, M. Mustafa, A. Albert, J.-L. Wu, C. Jiang, S. Esmaeilzadeh, K. Azizzadenesheli, R. Wang, A. Chattopadhyay, A. Singh, A. Manepalli, D. Chirila, R. Yu, R. Walters, B. White, H. Xiao, H. A. Tchelepi, P. Marcus, A. Anandkumar, P. Hassanzadeh, and Prabhat. Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379 (2194):20200093, 2021. doi: 10.1098/rsta.2020.0093.

E. J. Kendon, A. F. Prein, C. A. Senior, and A. Stirling. Challenges and outlook for convection-permitting climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2195):20190547, 2021. doi: 10.1098/rsta.2019.0547.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv*, 2017. doi: 10.48550/arXiv.1412.6980.

M. C. Knaus, M. Lechner, and A. Strittmatter. Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, 24(1):134–161, 2020. doi: 10.1093/ectj/utaa014.

R. D. Koster, P. A. Dirmeyer, Z. Guo, G. Bonan, E. Chan, P. Cox, C. T. Gordon, S. Kanae, E. Kowalczyk, D. Lawrence, P. Liu, C.-H. Lu, S. Malyshev, B. McAvaney, K. Mitchell, D. Mocko, T. Oki, K. Oleson, A. Pitman, Y. C. Sud, C. M. Taylor, D. Verseghy, R. Vasic, Y. Xue, and T. Yamada. Regions of Strong Coupling Between Soil Moisture and Precipitation. *Science*, 305(5687):1138–1140, 2004. doi: 10.1126/science.1100217.

F. Kratzert, D. Klotz, C. Brenner, K. Schulz, and M. Herrnegger. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11): 6005–6022, 2018. doi: 10.5194/hess-22-6005-2018.

H. Kraus. *Die Atmosphäre der Erde*. Springer Berlin Heidelberg, 2004. doi: 10.1007/3-540-35017-9.

M. Kretschmer, D. Coumou, J. F. Donges, and J. Runge. Using Causal Effect Networks to Analyze Different Arctic Drivers of Midlatitude Winter Circulation. *Journal of Climate*, 29(11):4069–4081, 2016. doi: 10.1175/jcli-d-15-0654.1.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

H. Lakkaraju, N. Arsov, and O. Bastani. Robust and Stable Black Box Explanations. In *37th International Conference on Machine Learning, PMLR*, volume 119, 2020. doi: 10.48550/arXiv.2011.06169.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6405—-6416. Curran Associates Inc., 2017. doi: 10.48550/arXiv.1612.01474.

S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10 (1), 2019. doi: 10.1038/s41467-019-08987-4.

P. R. Larraondo, L. J. Renzullo, I. Inza, and J. A. Lozano. A data-driven approach to precipitation parameterizations using convolutional encoder-decoder neural networks. *arXiv*, 2019. doi: 10.48550/arXiv.1903.10274.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient BackProp. In *Lecture Notes in Computer Science*, pages 9–48. Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-35289-8_3.

W. Lee, S. Kim, Y.-T. Lee, H.-W. Lee, and M. Choi. Deep neural networks for wild fire detection with unmanned aerial vehicle. In *2017 IEEE International Conference on Consumer Electronics (ICCE)*, pages 252–253, 2017. doi: 10.1109/ICCE.2017.7889305.

M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993. doi: 10.1016/s0893-6080(05)80131-5.

D. Leutwyler, A. Imamovic, and C. Schär. The Continental-Scale Soil-Moisture Precipitation Feedback in Europe with Parameterized and Explicit Convection. *Journal of Climate*, 34(13):1–56, 2021. doi: 10.1175/jcli-d-20-0415.1.

L. Li, W. Shangguan, Y. Deng, J. Mao, J. Pan, N. Wei, H. Yuan, S. Zhang, Y. Zhang, and Y. Dai. A Causal Inference Model Based on Random Forests to Identify the Effect of Soil Moisture on Precipitation. *Journal of Hydrometeorology*, 21(5):1115 – 1131, 2020. doi: 10.1175/JHM-D-19-0209.1.

W. Liu, Q. Zhang, C. Li, L. Xu, and W. Xiao. The influence of soil moisture on convective activity: a review. *Theoretical and Applied Climatology*, 149, 2022. doi: 10.1007/s00704-022-04046-z.

Y. Liu and L. Wu. Geological Disaster Recognition on Optical Remote Sensing Images Using Deep Learning. *Procedia Computer Science*, 91:566–575, 2016. doi: 10.1016/j.procs.2016.07.144.

Y. Liu, E. Racah, Prabhat, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. F. Wehner, and W. D. Collins. Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets. *CoRR*, abs/1605.01156, 2016. doi: 10.48550/arXiv.1605.01156.

J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965.

A. Loquercio, M. Segu, and D. Scaramuzza. A General Framework for Uncertainty Estimation in Deep Learning. *IEEE Robotics and Automation Letters*, 5(2):3153–3160, 2020. doi: 10.1109/lra.2020.2974682.

A. Massmann, P. Gentine, and J. Runge. Causal inference for process understanding in Earth sciences. *arXiv*, 2021. doi: 10.48550/arXiv.2105.00912.

R. Matthews. Storks Deliver Babies (p= 0.008). *Teaching Statistics*, 22(2):36–38, 2000. doi: 10.1111/1467-9639.00013.

A. McGovern, R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith. Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bulletin of the American Meteorological Society*, 100(11):2175–2199, 2019. doi: 10.1175/bams-d-18-0195.1.

F. Mesinger, G. DiMego, E. Kalnay, K. Mitchell, P. C. Shafran, W. Ebisuzaki, D. Jović, J. Woollen, E. Rogers, E. H. Berbery, M. B. Ek, Y. Fan, R. Grumbine, W. Higgins, H. Li, Y. Lin, G. Manikin, D. Parrish, and W. Shi. North American Regional Reanalysis. *Bulletin of the American Meteorological Society*, 87(3):343–360, 2006. doi: 10.1175/bams-87-3-343.

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

C. Molnar. *Interpretable Machine Learning*. 2022. `https://christophm.github.io/interpretable-ml-book`.

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

P. Nowack, J. Runge, V. Eyring, and J. D. Haigh. Causal networks for climate model evaluation and constrained projections. *Nature Communications*, 11(1), 2020. doi: 10.1038/s41467-020-15195-y.

A. Odena, V. Dumoulin, and C. Olah. Deconvolution and Checkerboard Artifacts. *Distill*, 2016. doi: 10.23915/distill.00003.

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

B. Pan, K. Hsu, A. AghaKouchak, and S. Sorooshian. Improving Precipitation Estimation Using Convolutional Neural Network. *Water Resources Research*, 55(3):2301–2321, 2019. doi: 10.1029/2018wr024090.

S. J. Pan and Q. Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/tkde.2009.191.

C. Papagiannopoulou, D. G. Miralles, S. Decubber, M. Demuzere, N. E. C. Verhoest, W. A. Dorigo, and W. Waegeman. A non-linear Granger-causality framework to investigate climate–vegetation dynamics. *Geoscientific Model Development*, 10(5):1945–1960, 2017. doi: 10.5194/gmd-10-1945-2017.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc., 2019. http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. doi: 10.48550/arXiv.1201.0490.

E. Perković, J. Textor, M. Kalisch, and M. H. Maathuis. Complete Graphical Characterization and Construction of Adjustment Sets in Markov Equivalence Classes of Ancestral Graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018. doi: 10.48550/arXiv.1606.06903.

J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc.: Series B (Statistical Methodology)*, 78 (5):947–1012, 2016. doi: 10.1111/rssb.12167.

E. Pichelli, E. Coppola, S. Sobolowski, N. Ban, F. Giorgi, P. Stocchi, A. Alias, D. Belušić, S. Berthou, C. Caillaud, R. M. Cardoso, S. Chan, O. B. Christensen, A. Dobler, H. de Vries, K. Goergen, E. J. Kendon, K. Keuler, G. Lenderink, T. Lorenz, A. N. Mishra, H.-J. Panitz, C. Schär, P. M. M. Soares, H. Truhetz, and J. Vergara-Temprado. The first multi-model ensemble of regional climate simulations at kilometer-scale resolution part 2: historical and future simulations of precipitation. *Climate Dynamics*, 56(11-12):3581–3602, 2021. doi: 10.1007/s00382-021-05657-4.

J. G. Powers, J. B. Klemp, W. C. Skamarock, C. A. Davis, J. Dudhia, D. O. Gill, J. L. Coen, D. J. Gochis, R. Ahmadov, S. E. Peckham, G. A. Grell, J. Michalakes, S. Trahan, S. G. Benjamin,

C. R. Alexander, G. J. Dimego, W. Wang, C. S. Schwartz, G. S. Romine, Z. Liu, C. Snyder, F. Chen, M. J. Barlage, W. Yu, and M. G. Duda. The Weather Research and Forecasting Model: Overview, System Efforts, and Future Directions. *Bulletin of the American Meteorological Society*, 98(8):1717–1737, 2017. doi: 10.1175/bams-d-15-00308.1.

E. Racah, C. Beckham, T. Maharaj, S. E. Kahou, Prabhat, and C. Pal. Extreme Weather: A Large-Scale Climate Dataset for Semi-Supervised Detection, Localization, and Understanding of Extreme Weather Events. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3405–3416. Curran Associates Inc., 2017. `https://dl.acm.org/doi/10.5555/3294996.3295099`.

M. Rahman, M. Sulis, and S. Kollet. The subsurface–land surface–atmosphere connection under convective conditions. *Advances in Water Resources*, 83:240–249, 2015. doi: 10.1016/j.advwatres.2015.06.003.

A. Rakhlin, A. Davydow, and S. Nikolenko. Land Cover Classification from Satellite Imagery with U-Net and Lovász-Softmax Loss. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 257–2574, 2018. doi: 10.1109/CVPRW.2018.00048.

H. Reichenbach. *The Direction of Time*. Dover Publications, 1956.

M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743): 195–204, 2019. doi: 10.1038/s41586-019-0912-1.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016. doi: 10.1145/2939672.2939778.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. doi: 10.48550/arXiv.1505.04597.

R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, 8:42200–42216, 2020. doi: 10.1109/access.2020.2976199.

A. Rotnitzky and E. Smucler. Efficient Adjustment Sets for Population Average Causal Treatment Effect Estimation in Graphical Models. *Journal of Machine Learning Research*, 21(188):1–86, 2020. doi: 10.48550/arXiv.1912.00306.

J. Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28:075310, 2018. doi: 10.1063/1.5025050.

J. Runge. Necessary and sufficient graphical conditions for optimal adjustment sets in causal graphical models with hidden variables. *arXiv*, 2021. doi: 10.48550/arXiv.2102.10324.

J. Runge, V. Petoukhov, and J. Kurths. Quantifying the Strength and Delay of Climatic Interactions: The Ambiguities of Cross Correlation and a Novel Measure Based on Graphical Models. *Journal of Climate*, 27(2):720–739, 2014. doi: 10.1175/jcli-d-13-00159.1.

J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, E. H. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1), 2019. doi: 10.1038/s41467-019-10105-3.

M. Sadeghi, P. Nguyen, K. Hsu, and S. Sorooshian. Improving near real-time precipitation estimation using a U-Net convolutional neural network and geographical information. *Environmental Modelling & Software*, 134:104856, 2020. doi: 10.1016/j.envsoft.2020.104856.

W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Muller. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*, 109 (3):247–278, 2021. doi: 10.1109/jproc.2021.3060483.

J. A. Santanello, P. A. Dirmeyer, C. R. Ferguson, K. L. Findell, A. B. Tawfik, A. Berg, M. Ek, P. Gentine, B. P. Guillod, C. van Heerwaarden, J. Roundy, and V. Wulfmeyer. Land–Atmosphere Interactions: The LoCo Perspective. *Bulletin of the American Meteorological Society*, 99(6): 1253–1272, 2018. doi: 10.1175/bams-d-17-0001.1.

C. Schär, D. Lüthi, U. Beyerle, and E. Heise. The Soil–Precipitation Feedback: A Process Study with a Regional Climate Model. *Journal of Climate*, 12(3):722–741, 1999. doi: 10.1175/1520-0442(1999)012<0722:tspfap>2.0.co;2.

L. Schneider, C. Barthlott, C. Hoose, and A. I. Barrett. Relative impact of aerosol, soil moisture, and orography perturbations on deep convection. *Atmospheric Chemistry and Physics*, 19(19): 12343–12359, 2019. doi: 10.5194/acp-19-12343-2019.

P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H. Luigs, A. Mahlein, and K. Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020. doi: 10.1038/s42256-020-0212-3.

D. L. Schumacher, J. Keune, C. C. van Heerwaarden, J. V.-G. de Arellano, A. J. Teuling, and D. G. Miralles. Amplification of mega-heatwaves through heat torrents fuelled by upwind drought. *Nature Geoscience*, 12(9):712–717, 2019. doi: 10.1038/s41561-019-0431-6.

K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8(1), 2017. doi: 10.1038/ncomms13890.

C. Schwingshackl, M. Hirschi, and S. I. Seneviratne. Quantifying Spatiotemporal Variations of Soil Moisture Control on Surface Energy Balance and Near-Surface Air Temperature. *Journal of Climate*, 30(18):7105–7124, 2017. doi: 10.1175/jcli-d-16-0727.1.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017. doi: 10.1109/iccv.2017.74.

S. I. Seneviratne, D. Lüthi, M. Litschi, and C. Schär. Land–atmosphere coupling and climate change in Europe. *Nature*, 443(7108):205–209, 2006. doi: 10.1038/nature05095.

S. I. Seneviratne, T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling. Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, 99(3-4):125–161, 2010. doi: 10.1016/j.earscirev.2010.02.004.

C. Shen. A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research*, 54(11):8558–8593, 2018. doi: 10.1029/2018wr022643.

C. Shi, D. M. Blei, and V. Veitch. Adapting Neural Networks for the Estimation of Treatment Effects. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2019. doi: 10.48550/arXiv.1906.02120.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227 – 244, 2000. doi: 10.1016/s0378-3758(00)00115-4.

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

P. Shrestha, M. Sulis, M. Masbou, S. Kollet, and C. Simmer. A Scale-Consistent Terrestrial Systems Modeling Platform Based on COSMO, CLM, and ParFlow. *Monthly Weather Review*, 142 (9):3466–3483, 2014. doi: 10.1175/MWR-D-14-00029.1.

B. Sibbald and M. Roland. Understanding controlled trials: Why are randomised controlled trials important? *BMJ*, 316(7126):201–201, 1998. doi: 10.1136/bmj.316.7126.201.

N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni. U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications. *IEEE Access*, 9:82031–82057, 2021. doi: 10.1109/ACCESS.2021.3086020.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv*, 2013. doi: 10.48550/arXiv.1312.6034.

A. Sinha, H. Namkoong, and J. C. Duchi. Certifiable Distributional Robustness with Principled Adversarial Training. In *International Conference on Learning Representations*, 2018. doi: 10.48550/arXiv.1710.10571.

W. Skamarock, J. Klemp, J. Dudhia, D. Gill, D. Barker, W. Wang, X.-Y. Huang, and M. Duda. A Description of the Advanced Research WRF Version 3. 2008. doi: 10.5065/D68S4MVH. Technical report.

J. V. Solórzano, J. F. Mas, Y. Gao, and J. A. Gallardo-Cruz. Land Use Land Cover Classification with U-Net: Advantages of Combining Sentinel-1 and Sentinel-2 Imagery. *Remote Sensing*, 13 (18), 2021. doi: 10.3390/rs13183600.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15 (1):1929–1958, 2014. http://jmlr.org/papers/v15/srivastava14a.html.

I. Sturm, S. Lapuschkin, W. Samek, and K. Müller. Interpretable deep neural networks for single-trial EEG classification. *Journal of Neuroscience Methods*, 274:141–145, 2016. doi: 10.1016/j.jneumeth.2016.10.008.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. doi: 10.48550/arXiv.1312.6199.

C. M. Taylor. Detecting soil moisture impacts on convective initiation in Europe. *Geophysical Research Letters*, 42(11):4631–4638, 2015. doi: 10.1002/2015gl064030.

C. M. Taylor, A. Gounou, F. Guichard, P. P. Harris, R. J. Ellis, F. Couvreux, and M. D. Kauwe. Frequency of Sahelian storm initiation enhanced over mesoscale soil-moisture patterns. *Nature Geoscience*, 4(7):430–433, 2011. doi: 10.1038/ngeo1173.

C. M. Taylor, R. A. M. de Jeu, F. Guichard, P. P. Harris, and W. A. Dorigo. Afternoon rain more likely over drier soils. *Nature*, 489(7416):423–426, 2012. doi: 10.1038/nature11377.

C. M. Taylor, C. E. Birch, D. J. Parker, N. Dixon, F. Guichard, G. Nikulin, and G. M. S. Lister. Modeling soil moisture-precipitation feedback in the Sahel: Importance of spatial scale versus convective parameterization. *Geophysical Research Letters*, 40(23):6213–6218, 2013. doi: 10.1002/2013gl058511.

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

T. Tesch, S. Kollet, and J. Garcke. Variant Approach for Identifying Spurious Relations That Deep Learning Models Learn - Software Code. 2021b. https://datapub.fz-juelich.de/slts/t_tesch/.

T. Tesch, S. Kollet, and J. Garcke. Causal deep learning models for studying the Earth system: soil moisture-precipitation coupling in ERA5 data across Europe - Software Code. *Zenodo*, 2022a. doi: 10.5281/ZENODO.6385040.

T. Tesch, S. Kollet, J. Garcke, E. Katragkou, and S. Kartsios. Data set of the manuscript “Opposite signs in local and nonlocal soil moisture-precipitation couplings across Europe”. *Juelich DATA*, 2022b. doi: 10.26165/JUELICH-DATA/YO3JCM.

T. Tesch, S. Kollet, and J. Garcke. Causal deep learning models for studying the Earth system. *Geoscientific Model Development*, 16(8):2149–2166, 2023a. doi: 10.5194/gmd-16-2149-2023. Highlight article.

T. Tesch, S. Kollet, J. Garcke, S. Kartsios, and E. Katragkou. Converse local and non-local soil-moisture–precipitation couplings across Europe. 2023b. Draft ready for submission to a scientific journal.

T. Tesch, S. Kollet, J. Garcke, S. Kartsios, and E. Katragkou. Opposite signs in local and nonlocal soil moisture-precipitation couplings across Europe - Software code. *Zenodo*, 2023c. doi: 10.5281/zenodo.7034237.

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

B. A. Toms, E. A. Barnes, and I. Ebert-Uphoff. Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability. *Journal of Advances in Modeling Earth Systems*, 12(9), 2020. doi: 10.1029/2019ms002002.

S. Tuttle and G. Salvucci. Empirical evidence of contrasting soil moisture-precipitation feedbacks across the United States. *Science*, 352(6287):825–828, 2016. doi: 10.1126/science.aaa7185.

S. E. Tuttle and G. D. Salvucci. Confounding factors in determining causal soil moisture-precipitation feedback. *Water Resources Research*, 53(7):5531–5544, 2017. doi: 10.1002/2016WR019869.

U.S. Geological Survey’s Water Science School. The Natural Water Cycle, 2019. `https://www.usgs.gov/media/files/natural-water-cycle-pdf`.

L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, M. Walczak, J. Pfrommer, A. Pick, R. Ramamurthy, J. Garcke, C. Bauckhage, and J. Schuecker. Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, 2021. doi: 10.1109/tkde.2021.3079836.

S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

B. Wallace and J. R. Minder. The impact of snow loss and soil moisture on convective precipitation over the Rocky Mountains under climate warming. *Climate Dynamics*, 56(9-10):2915–2939, 2021. doi: 10.1007/s00382-020-05622-7.

J. Wei and P. A. Dirmeyer. Sensitivity of land precipitation to surface evapotranspiration: a nonlocal perspective based on water vapor transport. *Geophysical Research Letters*, 46(21):12588–12597, 2019. doi: 10.1029/2019gl085613.

J. Welty and X. Zeng. Does Soil Moisture Affect Warm Season Precipitation Over the Southern Great Plains? *Geophysical Research Letters*, 45(15):7866–7873, 2018. doi: 10.1029/2018gl078598.

J. Witte, L. Henckel, M. H. Maathuis, and V. Didelez. On Efficient Adjustment in Causal Graphs. *Journal of Machine Learning Research*, 21(246):1–45, 2020. doi: 10.48550/arXiv.2002.06825.

J. Yoon, J. Jordon, and M. van der Schaar. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. In *International Conference on Learning Representations*, 2018. `https://openreview.net/forum?id=ByKWUeWA-`.

Q. Zhang and S. Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018. doi: 10.1631/fitee.1700808.

Z. Zhang, Q. Liu, and Y. Wang. Road Extraction by Deep Residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. doi: 10.1109/LGRS.2018.2802944.

X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017. doi: 10.1109/MGRS.2017.2762307.

Appendix

A. Variant approach for identifying spurious relations that deep learning models learn

Originally published in

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

A.1. Research article

frontiers in Water

ORIGINAL RESEARCH
published: 13 September 2021
doi: 10.3389/frwa.2021.745563

Variant Approach for Identifying Spurious Relations That Deep Learning Models Learn

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

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

A deep learning (DL) model learns a function relating a set of input variables with a set of target variables. While the representation of this function in form of the DL model often lacks interpretability, several interpretation methods exist that provide descriptions of the function (e.g., measures of feature importance). On the one hand, these descriptions may build trust in the model or reveal its limitations. On the other hand, they may lead to new scientific understanding. In any case, a description is only useful if one is able to identify if parts of it reflect spurious instead of causal relations (e.g., random associations in the training data instead of associations due to a physical process). However, this can be challenging even for experts because, in scientific tasks, causal relations between input and target variables are often unknown or extremely complex. Commonly, this challenge is addressed by training separate instances of the considered model on random samples of the training set and identifying differences between the obtained descriptions. Here, we demonstrate that this may not be sufficient and propose to additionally consider more general modifications of the prediction task. We refer to the proposed approach as variant approach and demonstrate its usefulness and its superiority over pure sampling approaches with two illustrative prediction tasks from hydrometeorology. While being conceptually simple, to our knowledge the approach has not been formalized and systematically evaluated before.

Keywords: interpretable deep learning, statistical model, machine learning, spurious correlation, causality, hydrometeorology, geoscience

OPEN ACCESS

Edited by:
Senlin Zhu,
Yangzhou University, China

Reviewed by:
Salim Heddam,
University of Skikda, Algeria
Rana Muhammad Adnan Ikram,
Hohai University, China

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

Specialty section:
This article was submitted to Water and Hydrocomplexity, a section of the journal Frontiers in Water

Received: *22 July 2021*
Accepted: *19 August 2021*
Published: *13 September 2021*

Citation:
Tesch T, Kollet S and Garcke J (2021) Variant Approach for Identifying Spurious Relations That Deep Learning Models Learn. Front. Water 3:745563. doi: 10.3389/frwa.2021.745563

1. INTRODUCTION

A deep learning (DL) model learns a function relating a set of input variables with a set of target variables. While DL models excel in terms of predictive performance, the representation of the learned function in form of the DL model (e.g., in form of a neural network) often lacks interpretability. To address this lack of interpretability, several interpretation methods have been developed (see e.g., Gilpin et al., 2018; Montavon et al., 2018; Zhang and Zhu, 2018; Molnar, 2019; Samek et al., 2021) providing descriptions of the learned function (e.g., measures of feature importance, FI). On the one hand, such descriptions can build trust in a model (Ribeiro et al., 2016) or reveal a model's limitations. Lapuschkin et al. (2019), for example, analyzed FI scores and found that their image classifier relied on a copyright tag on horse images. Similarly,

Schramowski et al. (2020) analyzed FI scores and found (and corrected) that their DL model classified sugar beet leaves as healthy or diseased while incorrectly focusing on areas outside of the leaves.

On the other hand, descriptions of the learned function can lead to new scientific understanding. Ham et al. (2019), for example, analyzed FI scores and identified a previously unreported precursor of the Central-Pacific El Niño type; Gagne et al. (2019) analyzed FI scores to gain a better understanding of the relations between environmental features and severe hail; McGovern et al. (2019) analyzed FI scores to gain a better understanding of the formation of tornadoes; and Toms et al. (2020) analyzed FI scores and identified regions related to the El Niño-Southern Oscillation (ENSO) and regions providing predictive capabilities for land surface temperatures at seasonal scales. Roscher et al. (2020) provide a general review of explainable machine learning for scientific insights in the natural sciences.

Whether descriptions of the function that a DL model learns are computed to build trust in the model, study the model's limitations, or gain new scientific understanding, it is important to identify if parts of a description reflect spurious instead of causal relations (e.g., random associations in the training data instead of associations due to a physical process). Examples for spurious relations are the above-mentioned copyright tag on horse images and the area outside of the classified sugar beet leaves. However, especially in prediction tasks involving physical, biological or chemical systems with several non-linearly interacting components, identifying spurious relations is challenging even for experts. Note that this does not only apply to the identification of spurious relations in descriptions of functions that DL models learn, but in general to the identification of spurious relations in descriptions of functions that any statistical model learns.

Commonly, this challenge is addressed by training separate instances of the considered model on random samples of the training set and aggregating or comparing the obtained descriptions. De Bin et al. (2015), for instance, compared subsampling and bootstrapping for the identification of relevant input variables in multivariable regression tasks. They applied a feature selection strategy repeatedly to samples of the original training set obtained by subsampling or bootstrapping, respectively, and identified relevant features by analyzing feature selection frequencies. As another example, Gagne et al. (2019) trained 30 instances of different statistical models on sampled training and test sets to take into account that the models' skills and the relations between input and target variables that the models learn might depend on the specific training and test set composition. Here, we propose to not only consider sampling, but also more general modifications of the original prediction task. We refer to this more general approach as variant approach. In the approach, separate instances of the considered statistical model (referred to as variant models) are trained 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 descriptions of the functions that original and variant models learn are compared and it is evaluated whether they reflect the assumed stability or specific variation, respectively, of causal relations. If this is not the case for some parts of the descriptions, these parts likely reflect spurious relations. The approach constitutes a generalization of sampling approaches in that sampling is one of many ways for modifying the original prediction task in order to obtain a variant task.

A similar concept to ours has, to the best of our knowledge, only been pursued systematically in a strict causality framework [for details on this framework see e.g., Pearl, 2009 or for a more methodological focus (Guo et al., 2020)]. Peters et al. (2016), for example, consider modifications of an original prediction task for which they require the conditional distribution $p(y|\vec{x}_S)$ of the target variable y given the complete set $\vec{x}_S$ of variables that directly cause y to remain stable. Exploiting this requirement, they aim to identify the subset S of (direct) causal predictors within all observed features. While this approach is conceptually related to the proposed variant approach, the latter does not require the strict causality framework but is applicable to any machine learning prediction task. Note that in our work the terms causal and spurious do not refer to an underlying causal graph or other concepts from the strict causality framework but should be interpreted with common sense: a pixel in an image, for instance, is causally related to the label "dog" if and only if it belongs to a dog in the image, and the value of a meteorological variable at a specific location and time is causally related to the value of a meteorological variable at another location and time if and only if one value influences the other via some physical process.

Other approaches in machine learning that consider modifications of an original prediction task predominantly aim to improve the predictive performance of a statistical model rather than to analyze the relations between input and target variables. Transfer learning (Pan and Yang, 2010), for instance, aims to extract knowledge from one or more source tasks to apply it to a target task, e.g., training a neural network first on a similar task before fine-tuning the weights on the target task. Adversarial training, as another example, optimizes the loss over a set of perturbations of the input (Goodfellow et al., 2015; Sinha et al., 2018) to become less susceptible to adversarial attacks (Szegedy et al., 2014), imperceptible changes to the input that can change the model's prediction. Traditional importance weighting (Shimodaira, 2000) or more recent methods (Lakkaraju et al., 2020), as further examples, shift the input distribution in order to perform better on a known or unknown test distribution.

In this work, we demonstrate the proposed variant approach with two illustrative prediction tasks from hydrometeorology. First, we predict the occurrence of rain at a target location, given geopotential fields at different pressure levels in a surrounding region. Second, we predict the water level at a location in a river, given the water level upstream and downstream 48 h earlier. As statistical models, we consider linear models and neural networks. After training a model on one of these tasks, we apply an interpretation method to obtain a description of the learned function. This description indicates the average importance of the different input locations for the predictions of the model. To identify if this importance reflects spurious instead of causal relations

between input and target variables, we apply the proposed variant approach.

The article is structured as follows: in section 2, we formalize the variant approach and define the two prediction tasks and variants thereof that illustrate the approach. Further, we introduce the statistical models and interpretation methods used in this work. Subsequently, we present and discuss the results obtained when training the statistical models on the considered prediction tasks and applying the variant approach. In section 4, we summarize our main findings and discuss perspectives for future research and applications of the variant approach.

2. MATERIALS AND METHODS

2.1. Variant Approach

During the training phase, a statistical model learns a function $f : \mathbb{R}^n \to \mathbb{R}^k$ relating an input space $X \subseteq \mathbb{R}^n$ with a target space $Y \subseteq \mathbb{R}^k$ given a training set $T = \{(\vec{x}_i, \vec{y}_i)\}_{i=1}^N$ with $\vec{x}_i \in X$, $\vec{y}_i \in Y$. As the representation of f in form of the statistical model (e.g., in form of a neural network) often lacks interpretability, several interpretation methods have been developed (see e.g., Gilpin et al., 2018; Montavon et al., 2018; Zhang and Zhu, 2018; Molnar, 2019; Samek et al., 2021). Most of these methods yield vector-valued descriptions $\vec{d} \in \mathbb{R}^d$ of f (e.g., measures of feature importance). These descriptions can be global or local, in the latter case not only depending on f but on a subset $X_\mathrm{d} \subset X$ as well. An example of a global description are the weights of a linear regression model. An example of a local description $\vec{d}(\vec{x})$ is the gradient of a neural network evaluated at a location $\vec{x} \in X$.

A description $\vec{d}$ reflects the relations between input and target variables that the statistical model learned. Whether the user aims to use $\vec{d}$ to build trust in the model, reveal the model's limitations, or gain new scientific understanding, it is important to identify if parts of the vector $\vec{d}$ reflect spurious instead of causal relations. In many cases, this is challenging even for experts. Therefore, we propose a variant approach. The approach consists of three steps. First, the original prediction task is modified in such a way that causal relations reflected in specific parts of $\vec{d}$ are assumed to either remain stable or vary in a specific way. We refer to the modified prediction task as variant task. Second, a separate instance of the considered statistical model (referred to as variant model) is trained on the variant task and a corresponding description $\vec{d}^v$ (referred to as variant description) of the function f^v that the variant model learns is computed. Third, original and variant descriptions are compared and it is evaluated whether the previously specified parts of original and variant descriptions reflect the assumed stability or specific variation, respectively, of causal relations. If this is not the case, the respective parts of the vector $\vec{d}$ or of the vector $\vec{d}^v$ reflect spurious relations.

Formalizing the approach, we define a variant task by an input space $X^v \subseteq \mathbb{R}^{n^v}$, a target space $Y^v \subseteq \mathbb{R}^{k^v}$, a training set $T^v = \{(\vec{x}_i^v, \vec{y}_i^v)\}_{i=1}^{N^v}$ with $\vec{x}_i^v \in X^v$, $\vec{y}_i^v \in Y^v$, an interpretation method (in most cases the same as for the original task) that provides a description $\vec{d}^v \in \mathbb{R}^{d^v}$ of the learned function $f^v : \mathbb{R}^{n^v} \to \mathbb{R}^{k^v}$, two sets of m boolean vectors $\vec{I}_j \in \{0,1\}^d$ and $\vec{I}_j^v \in \{0,1\}^{d^v}$, $j = 1, \ldots, m$, and m corresponding smooth (not necessarily symmetric) distance functions $dist_j : \mathbb{R}^d \times \mathbb{R}^{d^v} \to \mathbb{R}^{\geq 0}$, $j = 1, \ldots, m$. We denote by $\vec{d}(\vec{I}_j)$ [and analogously by $\vec{d}^v(\vec{I}_j^v)$] the restriction of $\vec{d}$ to the dimensions specified by the boolean vector $\vec{I}_j$ and refer to $\vec{d}(\vec{I}_j)$ as a *part* of $\vec{d}$. The distance function $dist_j$ incorporates the user's assumption about how the part $\vec{d}(\vec{I}_j)$ of $\vec{d}$ changes for the variant task if it reflects causal relations, and quantifies the deviation of this stability or specific variation, respectively. In other words, $dist_j$ computes a value $dist_j(\vec{d}, \vec{d}^v)$ which is 0 if $\vec{d}(\vec{I}_j)$ and $\vec{d}^v(\vec{I}_j^v)$ exhibit the assumed stability or systematic variation, respectively, of causal relations. In turn, the more they deviate from this assumed stability or specific variation, respectively, the larger the value $dist_j(\vec{d}, \vec{d}^v)$ should be.

Let us consider some examples of variant tasks. As already mentioned in the introduction, one way to modify the original prediction task in order to obtain a variant task is to consider a sampled training set, e.g., obtained by randomly sampling the original training set in the context of subsampling or bootstrapping (De Bin et al., 2015). In this case, we assume that all causal relations remain stable. Hence, we may choose to evaluate the dimensionwise distance between an original description $\vec{d} \in \mathbb{R}^d$ and the corresponding variant description $\vec{d}^v \in \mathbb{R}^d$ of the function f^v that a separate instance of the original model learns when trained on the sampled training set. Using the above formalism, this corresponds to defining the boolean vectors $(\vec{I}_j)_i = (\vec{I}_j^v)_i = \delta_{ji} \in \mathbb{R}^d$ (vectors with 0 components in all dimensions except from dimension j where the component is 1) and the distance functions $dist_j(\vec{d}, \vec{d}^v) = |\vec{d}_j - \vec{d}^v{}_j|$ for $j = 1, \ldots, m = d$. Now, $dist_j(\vec{d}, \vec{d}^v) \gg 0$ for some $j \in \{1, \ldots, d\}$ indicates that the part $\vec{d}(\vec{I}_j) = \vec{d}_j$ of the original description, or the part $\vec{d}^v(\vec{I}_j^v) = \vec{d}^v{}_j$ of the variant description, reflects spurious relations. Note that we can repeat the sampling procedure several times, leading to multiple variant tasks of the same type.

A second example for the definition of a variant task is to consider a modification of the input space. Later, for instance, we consider the task to predict a rain event at a target location given input variables in the 60 × 60 pixels neighborhood (see **Figure 1A**). As a variant task, we consider the input variables in the 80 × 80 pixels neighborhood instead. As original description $\vec{d} \in \mathbb{R}^{60\times 60}$, we consider a measure of the average importance of each pixel in the 60 × 60 pixels neighborhood for the predictions of the original model, and as variant description $\vec{d}^v \in \mathbb{R}^{80\times 80}$, we analogously measure the average importance of each pixel in the 80 × 80 pixels neighborhood for the predictions of the variant model. In this case, we assume that causal relations between pixels in the 60 × 60 pixels neighborhood and rain events at the target location remain stable when enlarging the considered neighborhood by 10 pixels on each side. Hence, we choose to evaluate the dimensionwise

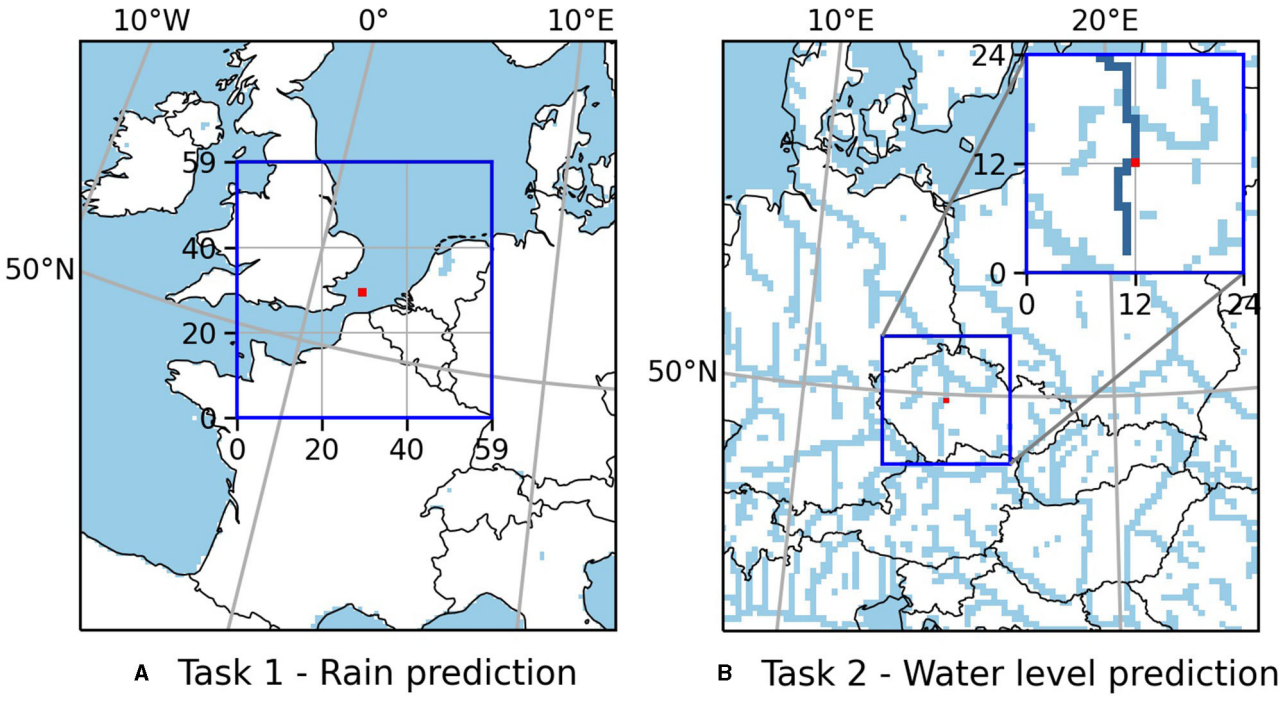

FIGURE 1 | Set up of the two original prediction tasks. **(A)** Predict whether the precipitation averaged over the red 2 × 2 pixels target patch in the center of the 60 × 60 pixels input region exceeds 1 mm in the next 3 h. **(B)** Predict the water level at the red pixel given the water level 48 h earlier at the red pixel and the pixels upstream and downstream marked dark blue in the inset. Light blue indicates pixels with ponded water at the land surface during the entire simulation period (rivers, lakes, …).

distance between the original description $\vec{d}$ and the central 60 × 60 pixels of the variant description $\vec{d}^v$. Using the above formalism, this corresponds to defining the boolean matrices $(\vec{I}_{j_1j_2})_{i_1i_2} = \delta_{j_1j_2,i_1i_2} \in \mathbb{R}^{60\times60}$ (matrices with 0 components in all dimensions except from dimension j_1j_2 where the component is 1), the boolean matrices $(\vec{I}^v_{j_1j_2})_{i_1i_2} = \delta_{j_1+10j_2+10,i_1i_2} \in \mathbb{R}^{80\times80}$ (10 corresponds to the offset between the neighborhoods for original and variant task, i.e., input index $(j_1 + 10, j_2 + 10)$ in the variant task corresponds to the same location as input index (j_1, j_2) in the original task) and the distance functions $dist_{j_1j_2}(\vec{d}, \vec{d}^v) = |\vec{d}_{j_1j_2} - \vec{d}^v{}_{j_1+10,j_2+10}|$ for $j_1, j_2 = 1, \ldots, 60$. Now, $dist_{j_1j_2}(\vec{d}, \vec{d}^v) \gg 0$ for some $j_1, j_2 \in \{1, \ldots, 60\}^2$ indicates that the part $\vec{d}(\vec{I}_{j_1j_2}) = \vec{d}_{j_1j_2}$ of the original description, or the part $\vec{d}^v(\vec{I}^v_{j_1j_2}) = \vec{d}^v{}_{j_1+10,j_2+10}$ of the variant description, reflects spurious relations. Note that for some statistical models, this type of variant task might require slight changes to the model architecture.

A third example for the definition of a variant task is to consider a modification of the target variable. Later, for instance, we predict the water level at a location in a river given the water level in some specified segment of the river (see **Figure 1B**). As a variant task, we consider the same segment of the river but shift the target location by τ pixels along the river (see **Figure 2B**). As original and variant descriptions $\vec{d}, \vec{d}^v \in \mathbb{R}^d$, we consider a measure of the average importance of each pixel in the specified river segment for the predictions of the original model and the variant model, respectively. In this case, we assume that causal relations are shifted along the river by the same distance as the target location is (i.e., by τ pixels). Hence, we choose to compute the dimensionwise distance between the original description $\vec{d}$ and the variant description $\vec{d}^v$ shifted by τ dimensions (i.e., we consider the distance $|\vec{d}_j - \vec{d}^v{}_{j+\tau}|$ for all j for that $j + \tau \in \{1, \ldots, d\}$). Using the above formalism, this corresponds to defining the boolean vectors $(\vec{I}_j)_i = (\vec{I}^v_j)_{i+\tau} = \delta_{ji}$ and the distance functions $dist_j(\vec{d}, \vec{d}^v) = |\vec{d}_j - \vec{d}^v{}_{j+\tau}|$ for all $j = 1, \ldots, d$ for that $j + \tau \in \{1, \ldots, d\}$. Now, $dist_j(\vec{d}, \vec{d}^v) \gg 0$ indicates that the part $\vec{d}(\vec{I}_j) = \vec{d}_j$ of the original description, or the part $\vec{d}^v(\vec{I}^v_j) = \vec{d}^v{}_{j+\tau}$ of the variant description, reflects spurious relations.

In this example, it might be more realistic to assume that causal relations are not shifted along the river by exactly τ pixels, but that the shift distance depends on the flow velocity and potentially further influences. The proposed formalism allows to take this into account by varying the definition of $\vec{I}_j$, $\vec{I}^v_j$ and $dist_j$. Suppose, for instance, that the flow velocity around the original target location is twice as high as around the shifted target location. In this case, we might assume that the sum of importance of the *two* pixels upstream of the original target location should be identical to the importance of the *single* pixel upstream of the shifted target location. Hence, we might decide to consider $(\vec{I}_j)_i = \delta_{ji} + \delta_{j-1,i}$, $(\vec{I}^v_j)_{i+\tau} = \delta_{ji}$ (as above), and $dist_j(\vec{d}, \vec{d}^v) = |(\vec{d}_j + \vec{d}_{j-1}) - \vec{d}^v{}_{j+\tau}|$, where the index j corresponds to the original target location. In this case, $dist_j(\vec{d}, \vec{d}^v) \gg 0$

indicates that the part $\vec{d}(\vec{I}_j)$ (corresponding to $\vec{d}_j$ *and* $\vec{d}_{j-1}$) of the original description, or the part $\vec{d}^v(\vec{I}_j^v) = \vec{d}^v{}_{j+\tau}$ of the variant description, reflects spurious relations.

In general, however, it is difficult to take variations of flow velocity and further influences into account when defining $\vec{I}_j$, $\vec{I}_j^v$ and $dist_j$. This is for example due to unavailable data on flow velocity and nonlinear behavior (e.g., that the sum of importance of the *two* pixels upstream of the original target location should be identical to the importance of the *single* pixel upstream of the shifted target location if the flow velocity in the respective river segment is twice as high, likely represents a too strong assumption on linearity). We will come back to this in the discussion of the results.

Let us return to the formal definition of the variant approach. The first step was to define a variant task. The second step consists of training a separate instance of the original model (a variant model) on this task and computing a variant description. The third step of the approach consists of comparing original and variant description and evaluating $dist_j(\vec{d}, \vec{d}^v) \gg 0$ for all $j = 1, \ldots, m$. If $dist_j(\vec{d}, \vec{d}^v) \gg 0$ for some $j \in \{1, \ldots, m\}$, the user infers that $\vec{d}(\vec{I}_j)$ or $\vec{d}^v(\vec{I}_j^v)$ reflects spurious relations. Note that the converse is not possible, i.e., if $dist_j(\vec{d}, \vec{d}^v) \approx 0$, the user cannot infer that $\vec{d}(\vec{I}_j)$ reflects causal relations (as it might be that both $\vec{d}(\vec{I}_j)$ and $\vec{d}^v(\vec{I}_j^v)$ reflect spurious relations). Note further that the specification of the condition $dist_j(\vec{d}, \vec{d}^v) \gg 0$ should in general take into account the specific original and variant task, the choice of the distance function $dist_j$, and the certainty of the assumed stability or systematic variation, respectively, of causal relations. Moreover, in case the user does not need a binary identification of parts of $\vec{d}$ that reflect spurious relations, it might be better not to consider the binary condition $dist_j(\vec{d}, \vec{d}^v) \gg 0$, but to consider raw values $dist_j(\vec{d}, \vec{d}^v)$, where higher distances indicate a higher probability that $\vec{d}(\vec{I}_j)$ or $\vec{d}^v(\vec{I}_j^v)$ reflects spurious relations.

For all variant tasks defined in this work, the expression $dist_j(\vec{d}, \vec{d}^v)$ corresponds to the relative distance between a single component $\vec{d}_{j_1}$ of an original description and a single component $\vec{d}^v{}_{j_2}$ of a corresponding variant description, i.e., it takes the form

$$dist_j(\vec{d}, \vec{d}^v) = \frac{|\vec{d}_{j_1} - \vec{d}^v{}_{j_2}|}{|\vec{d}_{j_1}| + |\vec{d}^v{}_{j_2}| + \varepsilon}, \tag{1}$$

with some regularization parameter $\varepsilon \geq 0$. By considering relative distances rather than absolute distances, we define, for instance, that $\vec{d}_{j_1} = 100$, $\vec{d}^v{}_{j_2} = 101$ agree better than $\vec{d}_{j_1} = 1$, $\vec{d}^v{}_{j_2} = 2$, or, in other words, in the latter case it is more likely that the value $\vec{d}_{j_1}$ or the value $\vec{d}^v{}_{j_2}$ reflects spurious relations. Further, an advantage of considering relative distances is that all distances lie between zero and one (when neglecting ε) which allows to apply a threshold $t \in (0, 1)$ to specify the condition $dist_j(\vec{d}, \vec{d}^v) \gg 0$ and to mark all parts $\vec{d}(I_j)$ of the original description as spurious for which $dist_j(\vec{d}, \vec{d}^v) > t$. In this study, we use $t = 0.5$ as threshold and $\varepsilon = 1e-3$ as regularization parameter. Choosing a smaller threshold, more values are marked as spurious (with all values marked as spurious for $t = 0$), and choosing a larger threshold, fewer values are marked as spurious (with no values marked as spurious for $t = 1$) by definition. For the examples considered below, $t = 0.5$ seems to be a good choice.

2.2. Illustrative Tasks

In this section, we define two prediction tasks and corresponding variant tasks that illustrate the proposed variant approach. We chose simplified tasks and global descriptions of the learned functions to be able to decide whether parts of the descriptions that the variant approach marks as spurious do indeed reflect spurious relations. The data underlying both tasks is 3-hourly data at 412 × 424 pixels over Europe. The data was obtained from a long-term (January 1996–August 2018), high-resolution (≈ 12.5 km) simulation (Furusho-Percot et al., 2019) performed with the Terrestrial Systems Modeling Platform (TSMP), a fully integrated groundwater-soil-vegetation-atmosphere modeling system (Gasper et al., 2014; Shrestha et al., 2014). Note that the statistical models and interpretation methods applied in this work are described in section 2.3.

2.2.1. Task 1 – Rain prediction

In the first example, we predict the occurrence of rain at a 2 × 2 pixels target patch, given the geopotential fields at 500, 850, and 1,000 hPa in the 60 × 60 pixels neighborhood (see **Figure 1A**). We model this as a classification task and define that rain occurred, if the precipitation averaged over the target patch exceeds 1 mm in the following 3 h. Previous works (Larraondo et al., 2019; Pan et al., 2019) have used CNNs to predict precipitation given geopotential fields to improve the parameterization of precipitation in numerical weather prediction models. Thus, apart from the simplifications of only one target location and a binary target, this is a realistic prediction task.

As statistical models, we consider a logistic regression model and two convolutional neural networks (CNNs) of different depth and complexity. As description of the function that the logistic regression model learns, we consider the absolute values of the model weights averaged over the pressure level axis. As descriptions of the functions that the CNNs learn, we consider saliency maps averaged over the pressure level axis and over all training samples. These descriptions can be seen as measures of the average importance of each pixel in the 60 × 60 pixels input region for the predictions of the models (for details see the respective sections below).

To identify whether parts of the descriptions reflect spurious relations that the models learned, we compute descriptions for variant models trained on three types of variant tasks. The first type (later referred to as sampling type) considers the same task, but a modified training set obtained by randomly sampling 70 % of the original training set without replacement. In this case, we assume that all causal relations remain stable. Hence, we compute the pixelwise distance between original and variant descriptions. We repeat the sampling procedure 10 times obtaining 10 variant tasks of this type. The second type of variant tasks (later referred to as size type) considers the same task but the input variables in the 80 × 80 pixels neighborhood of the target patch. In

this case, we assume that causal relations between pixels in the 60×60 pixels neighborhood and rain events at the target patch remain stable when enlarging the considered neighborhood by 10 pixels on each side. Hence, we compute the pixelwise distance between the original descriptions and the central 60×60 pixels of the variant descriptions. The third type of variant task (later referred to as location type) considers the same task but for eight different target patches obtained by moving the original target patch by five pixels to the left or right, and up or down. The input regions are shifted accordingly (see **Figure 2A**). In this case, we assume again that all causal relations remain stable. Hence, we again compute the pixelwise distance between original and variant descriptions.

Note that to compute the variant descriptions for the functions that separate instances of the CNNs learn when trained for different target locations, we average the saliency maps over all training samples from the *original* task. This is because the distribution $p(\vec{x})$ of geopotential fields differs at different locations. Thus, if we averaged the saliency maps for a variant CNN over all training samples from a variant task, the obtained variant description would differ from the original description even if original and variant models learned the exact same function relating geopotential fields and rain events.

We obtained the geopotential fields and precipitation data from the aforementioned simulation. We selected the geopotential fields in the considered input regions and created the binary rain event time series for the corresponding target patches. Next, we split the time series using the first 56,000 time steps as training candidates and the last 10,183 time steps as validation candidates. Finally, training and validation sets were obtained by selecting all time steps followed by a rain event at the considered target patch and an equal amount of randomly chosen additional time steps for non-rain events from the training and validation candidates, respectively. This resulted in balanced training and validation sets of a total of approximately 10,000 time steps for each target patch. Handling strongly unbalanced data sets as it would be necessary without such a selection of time steps is out of scope for this work.

2.2.2. Task 2 – Water Level Prediction

As a second example, we predict the water level at a location in a river, given the water level in a specific segment of the river 48 h earlier (see **Figure 1B**).

As statistical models, we consider a linear regression model and a multilayer perceptron (MLP). As description of the function that the linear regression model learns, we consider as in Task 1 the absolute values of the model weights. For the MLP, we consider again the saliency maps averaged over all training samples. Analogously to Task 1, these descriptions can be seen as measures of the average importance of each pixel in the considered river segment for the predictions of the models (for details see the respective sections below).

To identify whether parts of the descriptions reflect spurious relations that the models learned, we compute descriptions for variant models trained on two types of variant tasks. The first type (later referred to as sampling type) considers the same task, but a modified training set obtained by randomly sampling 70 % of the original training set without replacement. In this case, we assume that all causal relations remain stable. Hence, we compute the pixelwise distance between original and variant descriptions. We repeat the sampling procedure 10 times obtaining 10 variant tasks of this type. The second type of variant tasks (later referred to as location type) considers the same river segment as input, but target locations closely upstream and downstream of the original target location (see **Figure 2B**). In this case, we assume that causal relations are shifted along the river by the same distance as the target location is. Hence, we compute the pixelwise distance between the original description $\vec{d}$ and the variant description $\vec{d}^v$ shifted by τ pixels, where τ is the number of pixels that the target location was shifted (i.e., we consider the distance $|\vec{d}_j - \vec{d}^v{}_{j+\tau}|$ for all j for that $j + \tau \in \{1, \ldots, d\}$).

We obtained the water level data from the aforementioned simulation. In contrast to Task 1, this task is not a classification but a regression task; discarding time steps to obtain a balanced data set is not necessary. Hence, we use water level data for all 64,240 3-hourly time steps between January 1996 and December 2017. We randomly selected the years 1997, 2004, 2008, and 2015 as test data, covering the whole period of time, and use the remaining years to train the models.

2.3. Statistical Models and Descriptions

In this section, we present the statistical models used in this study. Further, we describe saliency maps, the interpretation method applied to obtain descriptions of the functions that the neural networks (MLP and CNNs) learn. Note that for the considered examples, layerwise relevance propagation (LRP) and Grad-CAM give very similar results to saliency maps. The section is ordered with respect to the complexity of the described methods from simple to complex.

2.3.1. Linear Regression

Given training samples $(\vec{x}_i, y_i)_{i=1}^n$ with $\vec{x}_i \in \mathbb{R}^N$, $y_i \in \mathbb{R}$, a linear regression model learns a function $f : \mathbb{R}^N \to \mathbb{R}$ of the form

$$f(\vec{x}) = \beta_0 + \vec{x}^T \cdot \vec{\beta}, \tag{2}$$

where $\vec{\beta} = (\beta_0, \vec{\beta}) = (\beta_0, \beta_1, \ldots, \beta_N) \in \mathbb{R}^{N+1}$ are the weights of the model. Those weights are obtained by minimizing the squared error on the training set

$$\sum_{i=1}^{n} (f(\vec{x}_i) - y_i)^2. \tag{3}$$

Optionally, a regularization term can be added to the objective. We calculate the minimizing weights $\vec{\beta}$ using the implementation of scikit-learn (Pedregosa et al., 2011). In our case, the inputs $\vec{x}_i$ are elements of $\mathbb{R}^{30}$ representing the water level at the 30 pixels in the considered river segment (see **Figure 1B**) and the targets $y_i \in \mathbb{R}$ represent the water level at the target pixel 48 h later.

As description of the function that a linear regression model learned, we consider the absolute values of the weights $\vec{\beta}$. This can be seen as a measure of the average importance of each pixel in the river segment for the predictions of the model (Molnar, 2019).

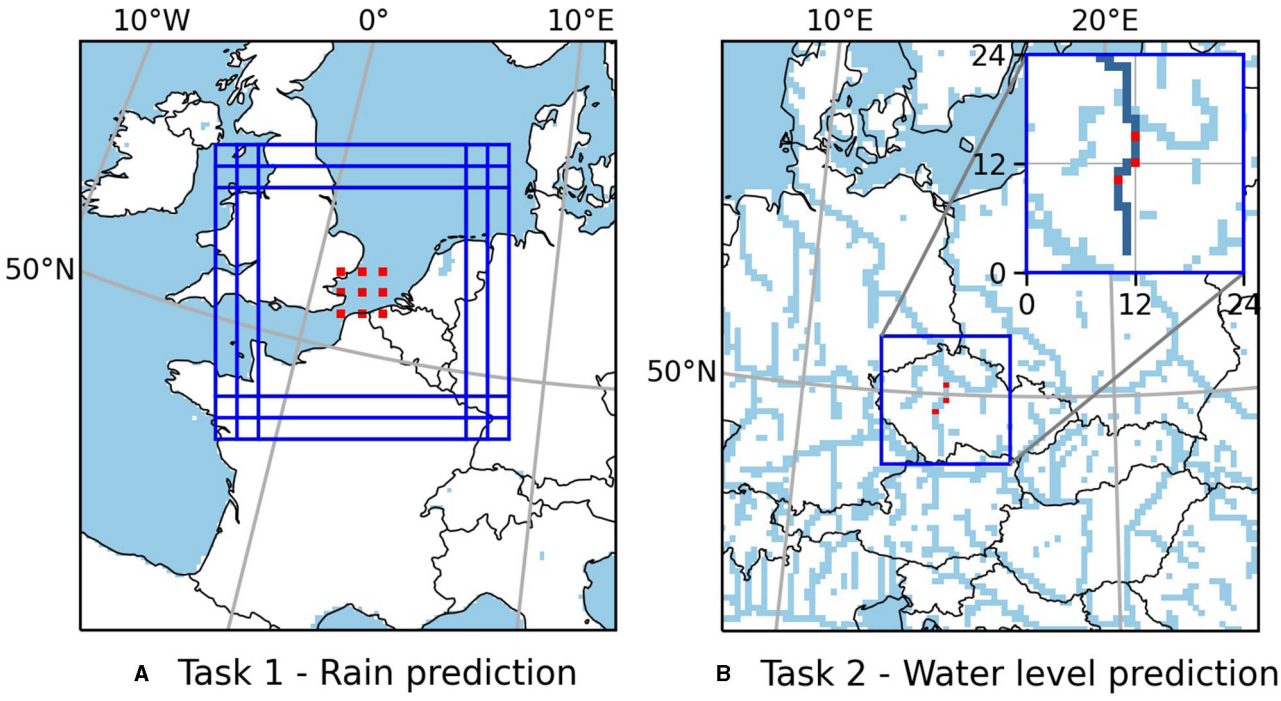

FIGURE 2 | Location variant tasks. **(A)** Original target patch (center) with its input region and eight additional target patches and their (overlapping) input regions. **(B)** Original target location (center) and two additional target locations closely upstream and downstream.

2.3.2. Logistic Regression

Given the task to predict a binary target $y \in \{0, 1\}$ from an input $\vec{x} \in \mathbb{R}^N$, a logistic regression model yields

$$P(y = 1|\vec{x}, \vec{\beta}) = \frac{1}{1 + \exp(-(\beta_0 + \vec{x}^T \cdot \vec{\beta}))}, \quad (4)$$

where $\vec{\beta} = (\beta_0, \vec{\beta}) = (\beta_0, \beta_1, \ldots, \beta_N) \in \mathbb{R}^{N+1}$ are the weights of the model. These weights are obtained by minimizing the function

$$-\prod_{i=1}^{n} P(y_i = 1|\vec{x}_i, \vec{\beta})^{y_i} \cdot (1 - P(y_i = 1|\vec{x}_i, \vec{\beta}))^{1-y_i} + \lambda R(\vec{\beta}) \quad (5)$$

with respect to $\vec{\beta}$. Here, $(\vec{x}_i, y_i)_{i=1}^n$ are training samples with $\vec{x}_i \in \mathbb{R}^N, y_i \in \{0, 1\}$, and $\lambda R(\vec{\beta})$ is a regularization term. The product represents the probability with that – according to the logistic regression model with weights $\vec{\beta}$ – the targets y_i are observed given the input samples $\vec{x}_i$. Thus, minimizing the negative product with respect to $\vec{\beta}$ corresponds to finding the $\vec{\beta}$ for that the highest probability is assigned to observing the targets y_i given the inputs $\vec{x}_i$ from the training set. We use scikit-learn (Pedregosa et al., 2011) (solver "liblinear") to approximate the minimizing weights $\vec{\beta}$. In our case, the inputs $\vec{x}_i$ are the geopotential fields at 500, 850 and 1000 hPa flattened to vectors in $\mathbb{R}^{3 \cdot 60 \cdot 60}$ and the targets $y_i \in \{0, 1\}$ represent whether a rain event took place or not.

As description of the function that a logistic regression model learned, we consider the weights $\vec{\beta}$. We reshape the vector $\vec{\beta}$ to the shape of the original input, $3 \times 60 \times 60$, take the absolute value and build an average over the first (pressure level) axis. This can be seen as a measure of the average importance of each pixel in the 60×60 pixels input region for the predictions of the model (Molnar, 2019).

2.3.3. Multilayer Perceptron

Multilayer Perceptrons (MLPs), also referred to as fully-connected neural networks, are feedforward artificial neural networks. They are composed of one or more hidden layers and an output layer. Each layer comprises several neurons. Each neuron in the first hidden layer builds a weighted sum of all input variables, while each neuron in the subsequent layers builds a weighted sum of the outputs of the neurons in the respective previous layer. In case of a neuron in a hidden layer, the sum is passed through a nonlinear activation function and forms the input to the next layer. In case of a neuron in the output layer, the sum is optionally passed through a nonlinear activation function and forms the output of the neural network. The weights of the MLP are learned by minimizing a loss function on training samples $(\vec{x}_i, \vec{y}_i)_{i=1}^n$, $\vec{x}_i \in \mathbb{R}^N, \vec{y}_i \in \mathbb{R}^K$, using backpropagation (LeCun et al., 2012).

In our case, the inputs to the MLP are elements $\vec{x}$ of $\mathbb{R}^{30}$ representing the water level at the 30 pixels in the considered river segment (see **Figure 1B**). The targets $y_i \in \mathbb{R}$ represent the water level at the target pixel 48 h later. Section 2.3.5 describes how

we obtained a description of the function that the MLP learned. The network and training of the MLP were implemented using the deep learning library Pytorch (Paszke et al., 2019). A detailed description of the used architecture and training procedure can be found in the **Supplementary Material**.

2.3.4. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are frequently employed DL models designed to process stacks of multiple arrays containing spatially structured data. This can, for example, be a stack of 2-dimensional arrays for an RGB image ($\vec{x}_i \in \mathbb{R}^{3\times\text{height}\times\text{width}}$) or, as in our case, a stack of 2-dimensional geopotential fields at different pressure levels in the atmosphere ($\vec{x}_i \in \mathbb{R}^{3\times 60\times 60}$). Typically, a CNN consists of three types of layers: convolutional layers, pooling layers and fully-connected layers. In the following short review of the typical CNN layers, we consider the case of one or multiple 2-dimensional input arrays. A generalization of the concepts to N-dimensional input arrays is straightforward.

The input to a convolutional layer is a stack of c_{in} 2-dimensional arrays and its output is a stack of c_{out} 2-dimensional arrays. The convolutional layer is characterized by c_{out} kernels, which are 3-dimensional tensors of shape $c_{\text{in}} \times k \times k$, where the kernel size k is usually between 1 and 7. The output of the layer are the c_{out} 2-dimensional arrays obtained by convolving the input with each kernel along the last two dimensions. Usually, a convolutional layer is directly followed by a nonlinear activation function which is applied elementwise to the layer's output. In contrast to a fully-connected layer, a convolutional layer preserves the spatial structure of the input: only neurons in a neighborhood defined by the kernel size influence the output of a specific neuron.

As for convolutional layers, the input to a pooling layer is a stack of c_{in} 2-dimensional arrays of shape $n \times m$. Pooling layers reduce the dimensionality of the 2-dimensional arrays creating invariances to small shifts and distortions. A typical form of pooling is max-pooling with a kernel size of two. This reduces the resolution along both axes of each of the c_{in} 2-dimensional arrays by a factor of two, picking always the maximum value of a 2×2 patch of the original array. Thus, the output of this pooling layer is a stack of $c_{\text{out}} = c_{\text{in}}$ 2-dimensional arrays of shape $\frac{n}{2} \times \frac{m}{2}$.

After several alternating convolutional and pooling layers which extract features of increasing complexity, the resulting c 2-dimensional arrays are flattened into a single vector and one or more fully-connected layers, as described for the MLP, follow. The weights for the kernels in the convolutional layers and the fully-connected layers are learned by minimizing a loss function on training samples $(\vec{x}_i, \vec{y}_i)_{i=1}^n$, $\vec{x}_i \in \mathbb{R}^N, \vec{y}_i \in \mathbb{R}^K$, using backpropagation (LeCun et al., 2012). To prevent CNNs from overfitting, dropout regularization (Srivastava et al., 2014) and batch normalization (Ioffe and Szegedy, 2015) are commonly employed techniques.

In our case, the inputs $\vec{x}_i$ are the geopotential fields at 500, 850 and 1000 hPa, $\vec{x}_i \in \mathbb{R}^{3\times 60\times 60}$. The targets $y_i \in \{0, 1\}$ represent whether a rain event took place or not. We consider two convolutional neural networks of different depth and complexity. CNN1 is a shallow CNN with only two convolutional layers followed by a single fully-connected layer. CNN2 is a commonly employed, much deeper CNN architecture called resnet18 (He et al., 2016) for which the last fully-connected layer was adapted to have only two output neurons to fit our binary prediction task. Section 2.3.5 describes how we obtained descriptions of the functions that the CNNs learned. The networks and training were implemented using the deep learning library Pytorch (Paszke et al., 2019). A detailed description of the used CNN architectures and training procedure can be found in the **Supplementary Material**.

2.3.5. Saliency Maps

A common subgroup of interpretation methods providing descriptions of the functions that neural networks (NNs) learn, are methods that assign an importance to each dimension of individual input samples $\vec{x} \in \mathbb{R}^N$ (local feature importance scores), see e.g., Samek et al., 2021. Among the most employed and well-known methods for that purpose are saliency maps (Simonyan et al., 2014), layerwise relevance propagation (LRP) (Bach et al., 2015) and Grad-CAM (Selvaraju et al., 2017). In the examples presented in this work, all three methods yield similar results. Therefore and for the sake of brevity, we focus on saliency maps (although e.g., Montavon et al., 2018 argue that saliency maps provide a bad measure of feature importance because they indicate how the prediction of a model changes when the value of a feature is changed, rather than indicating what makes the model make a prediction).

Note that in contrast to the weights of linear and logistic regression models, saliency maps are local descriptions of the learned functions, i.e., the importance assigned to an input dimension (in our case an input pixel) depends on the input sample $\vec{x}$. To get a global description of the learned function and a measure of the average importance of each input pixel, we average the saliency maps over all training samples.

In the rain prediction task, the NN defines an (almost everywhere) differentiable function f that maps geopotential fields $\vec{x} \in \mathbb{R}^{3\times 60\times 60}$ to probabilities $f(\vec{x}) = y \in (0, 1)$ that a rain event occurs. The partial derivative

$$w_{cij}(\vec{x}) = \frac{\partial f}{\partial x_{cij}}(\vec{x}), \quad c = 1, 2, 3, \quad i, j = 1, \ldots, 60 \tag{6}$$

indicates how a small perturbation of the c-th geopotential field at pixel (i, j) affects the prediction of the NN. The saliency map

$$M_{ij}(\vec{x}) = \frac{1}{3}\sum_{c=1}^{3} |w_{cij}(\vec{x})|, \quad i, j = 1, \ldots, 60 \tag{7}$$

considers the absolute value of the partial derivatives averaged over the pressure level axis to obtain for each pixel in the 60×60 pixels input region a measure of its importance for the model's prediction for sample $\vec{x}$.

In the water level prediction task, the neural network maps water levels $\vec{x} \in \mathbb{R}^{30}$ to a water level prediction $f(\vec{x}) = y \in \mathbb{R}$. The saliency map

$$M_i(\vec{x}) = |w_i(\vec{x})| = \left|\frac{\partial f}{\partial x_i}(\vec{x})\right|, \quad i = 1, \ldots, 30 \tag{8}$$

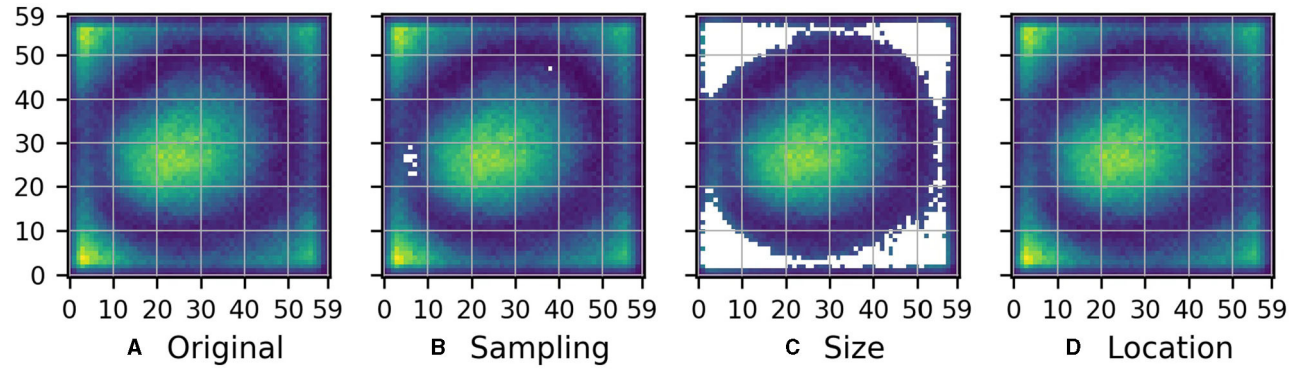

FIGURE 3 | (A) Description $\vec{d}$ of the function that CNN1 learned when it was trained on the original rain prediction task (see **Figure 1A**). The description is a measure of the average importance of each pixel in the 60 × 60 pixels input region for the predictions of the model. Yellow color indicates high and blue color low importance. **(B–D)** As **(A)**, but pixels for which the relative distance between the original description $\vec{d}$ and one of the variant descriptions $\vec{d^{v_i}}$ obtained for the sampling, size, and location variant tasks, respectively, exceeds the threshold of $t = 0.5$, are masked.

provides for each pixel in the considered river segment a measure of its importance for the model's prediction for sample $\vec{x}$.

3. RESULTS AND DISCUSSION

3.1. Task 1 – Rain Prediction

Figure 3A shows the description $\vec{d}$ of the function that CNN1 learned when it was trained on the original rain prediction task. Remember that the considered description is a measure of the average importance of each pixel in the 60 × 60 pixels input region for the predictions of the model. Our objective is to apply the variant approach to identify parts of the description that reflect spurious relations. To that purpose, we defined several variant tasks above. As a next step, we computed the corresponding variant descriptions, i.e., the descriptions of the functions that separate instances of CNN1 learned when trained on these variant tasks. For illustration, **Figure 4** shows the original description (center, same as **Figure 3A**) and the variant descriptions $\vec{d^{v_i}}, i = 1, \ldots, 8$, obtained for the eight location variant tasks (see **Figure 2A**).

For each of these variant descriptions $\vec{d^{v_i}} \in \mathbb{R}^{60\times60}, i = 1, \ldots, 8$, we evaluated the pixelwise relative distance to the original description $\vec{d} \in \mathbb{R}^{60\times60}$ (see Equation 1), and masked all pixels of the original description $\vec{d}$ for which this distance exceeds the threshold of $t = 0.5$ for any $\vec{d^{v_i}}$. The resulting masked version of $\vec{d}$ is shown in **Figure 3D**. Note that in this case, there is no pixel for which the relative distance between original description and any of the variant descriptions exceeds 0.5, hence **Figure 3D** is identical to **Figure 3A**. Analogously to **Figures 3B,D** shows the masked version of $\vec{d}$ obtained when masking all pixels for which the pixelwise relative distance between $\vec{d}$ and one of the variant descriptions $\vec{d^{v_i}}$ obtained for the sampling variant tasks exceeds 0.5. We observe that some pixels in the west of the inner area of importance are masked, indicating that the inner area of importance might actually extend further to the west.

Figure 3C shows the masked version of $\vec{d}$ obtained when masking all pixels for which the pixelwise relative distance between $\vec{d}$ and the central 60 × 60 pixels of the variant description $\vec{d^{v_i}}$ obtained for the size variant task exceeds 0.5. Notably, all the boundary pixels with high values in **Figure 3A** are masked, indicating that these values likely reflect spurious relations.

Figure 5 shows the same as **Figure 3** but for CNN2. Only few pixels are masked for the sampling and location variant tasks. However, the mask obtained for the size variant task indicates that the checkerboard pattern in the original description $\vec{d}$, which is shown in **Figure 5A**, likely reflects spurious relations. Note that this checkerboard pattern is indeed a known artifact of strided convolutions and max-pooling layers used in CNN2 (Odena et al., 2016).

Figure 6 shows the same as **Figures 3** and **5** but for the logistic regression model. For the sampling variant tasks, large parts of the original description $\vec{d}$ are masked. This indicates that these parts likely reflect spurious relations. For the size variant task, on the other side, only few pixels are masked. Lastly, for the location variant tasks, nearly all pixels are masked. This indicates that the original description $\vec{d}$ shown in **Figure 6A** likely reflects spurious relations only.

For this task, we know that the physical importance of a pixel averaged over a long time period decreases with the pixel's distance to the central target patch. Further, due to the predominantly westerly winds, the average physical importance of pixels is slightly shifted to the west. Given this knowledge, we can confirm that the variant approach successfully identified all pixels in **Figures 3A**, **5A**, **6A** which reflect spurious relations. Note that the sampling approach alone (see **Figures 3B**, **5B**, **6B**), which is the commonly applied method, is not sufficient to identify all pixels reflecting spurious relations.

Note further that the examples emphasize once again the following: even if parts of a description are not indicated as spurious by any considered variant task, we cannot conclude that they reflect causal relations. Imagine, for instance, that

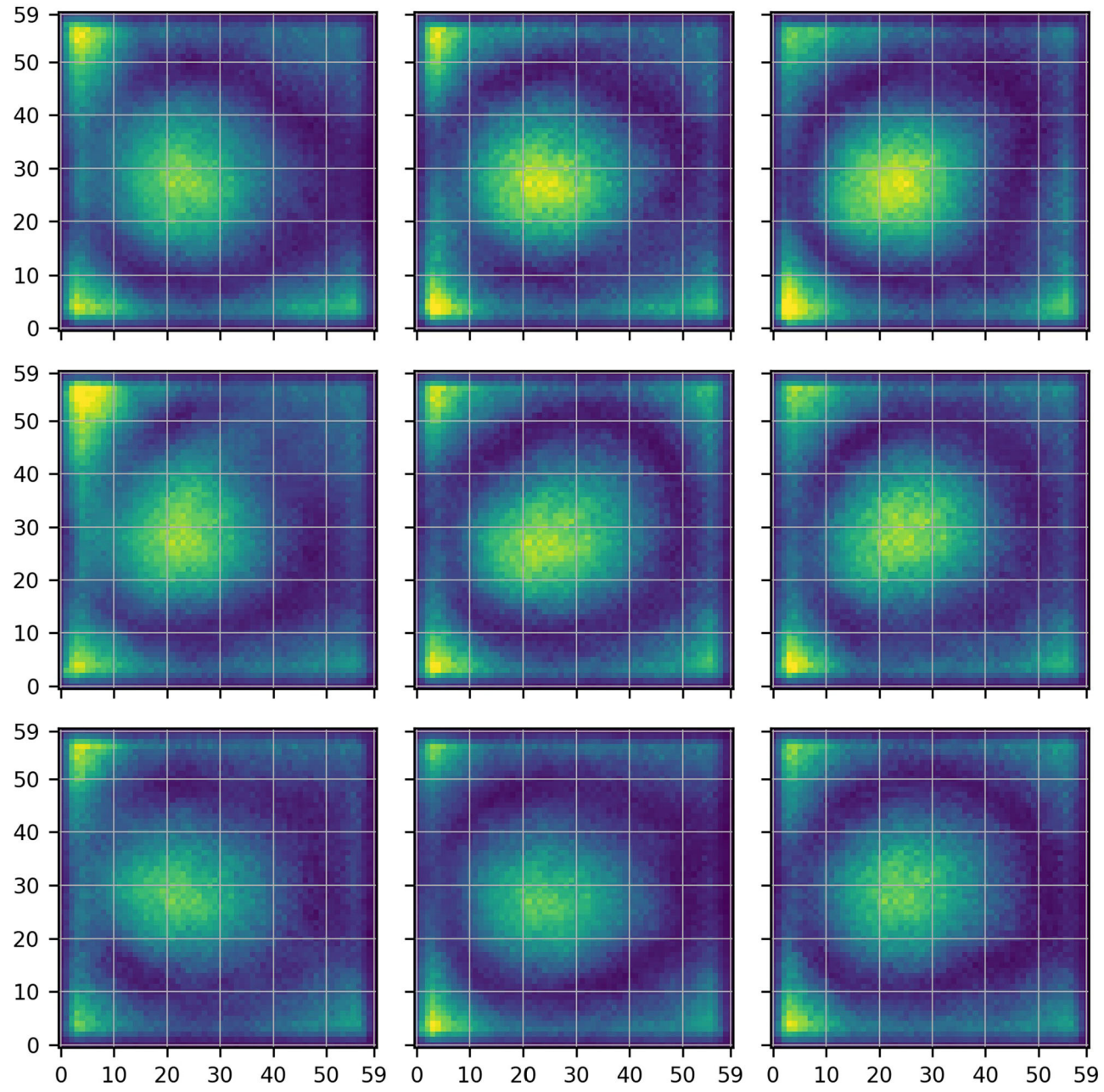

FIGURE 4 | Descriptions obtained when training separate instances of CNN1 for the nine different locations depicted in **Figure 2A**. Each description is a measure of the average importance of each pixel in the 60 × 60 pixels input region for the predictions of the respective instance of CNN1. Yellow color indicates high and blue color low importance. The central location is the original target location, hence the central description identical to **Figure 3A**.

we had only considered the size variant task. For this variant task and the logistic regression model, only a small number of pixels is masked although **Figure 6A** seems to exclusively reflect spurious relations. Hence, variant tasks can only indicate parts of an original description as likely reflecting spurious relations and do not allow for any direct inference about other parts of the description. Nevertheless, this can be useful already.

3.2. Task 2 – Water Level Prediction

Figure 7A shows the description $\vec{d}$ of the function that the MLP learned when it was trained on the original water level prediction task. Remember that the considered description is a measure of the average importance of each pixel in the considered river segment for the predictions of the model. Our objective is to apply the variant approach to identify parts of the description that reflect spurious relations. To that purpose we computed

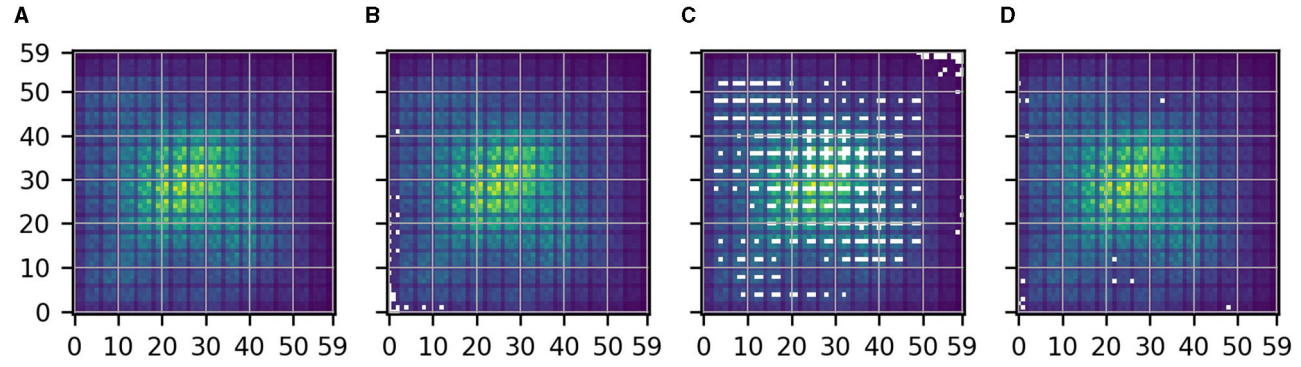

FIGURE 5 | Same as **Figure 3** but for CNN2. **(A)** Original. **(B)** Sampling. **(C)** Size. **(D)** Location.

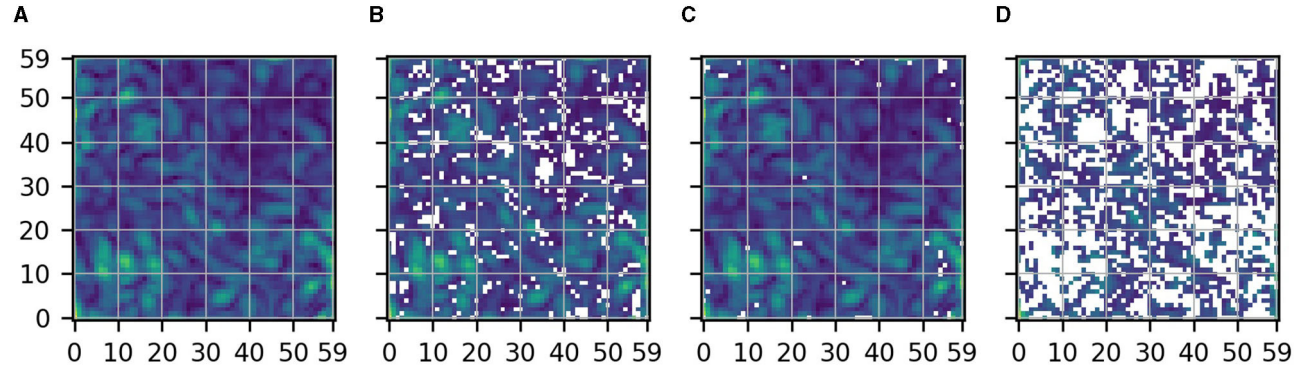

FIGURE 6 | Same as **Figure 3** but for the logistic regression model. **(A)** Original. **(B)** Sampling. **(C)** Size. **(D)** Location.

the variant descriptions $\vec{d}^{v_i}$ for all sampling and location variant tasks, and masked all pixels of $\vec{d}$ for which the relative distance between the original description $\vec{d}$ and one of the (shifted) variant descriptions exceeds the threshold of $t = 0.5$. The resulting masked versions of **Figure 7A** are shown in **Figures 7B, C**.

For this task, we know that the development of the water level at the target location depends only on the water level closely upstream and downstream. Hence, **Figure 7A** is [apart from the moderately high importance of pixel (11,17)] close to our understanding of the physical importance of the considered pixels. Nevertheless, especially in **Figure 7C**, many of the pixels further upstream and downstream of the target location are masked, i.e., the variant approach indicates (mistakenly) that the low feature importance of these pixels likely reflects spurious relations. We suspect that this happened because we considered relative rather than absolute distances between original and variant descriptions (see Equation 1), which can cause two small values to have a large distance which in turn causes the corresponding pixel to be mistakenly masked as spurious. Apart from pixels with low feature importance, also pixel (11,11) closely upstream of the original target location seems to be mistakenly masked as spurious in **Figure 7C**. We suspect that this is due to our assumption that causal relations are shifted along the river by the exact same number of pixels as the target location is. While this assumption enables us to simply consider pixelwise relative distances between original description $\vec{d}$ and shifted variant descriptions $\vec{d}^{v_i}$ (see section 2), it might be overly simplified as for example the flow velocity at different locations in the river might differ, and the river might cross some pixels diagonally and others straight.

Here, a visual assessment of the individual variant descriptions seems to be superior to the formal evaluation of distances performed for **Figure 7C** because it allows a softer comparison between original and variant descriptions $\vec{d}$ and $\vec{d}^{v_i}$. Indeed, upon visual assessment of the location variant descriptions depicted in **Figure 8**, and with the assumption in mind that causal relations *approximately* reflect the shift of the target location, the only pixel in **Figure 7A** that we would mark as potentially reflecting spurious relations, is pixel (11,17).

Figures 9, **10** show the same as **Figures 7**, **8** but for the linear regression model. In this case, the formal evaluation of distances between original and location variant descriptions performed for **Figure 9C** indicates that **Figure 9A** reflects spurious relations at nearly all pixels except from the target location and the neighboring pixel upstream. In this case, the formal evaluation

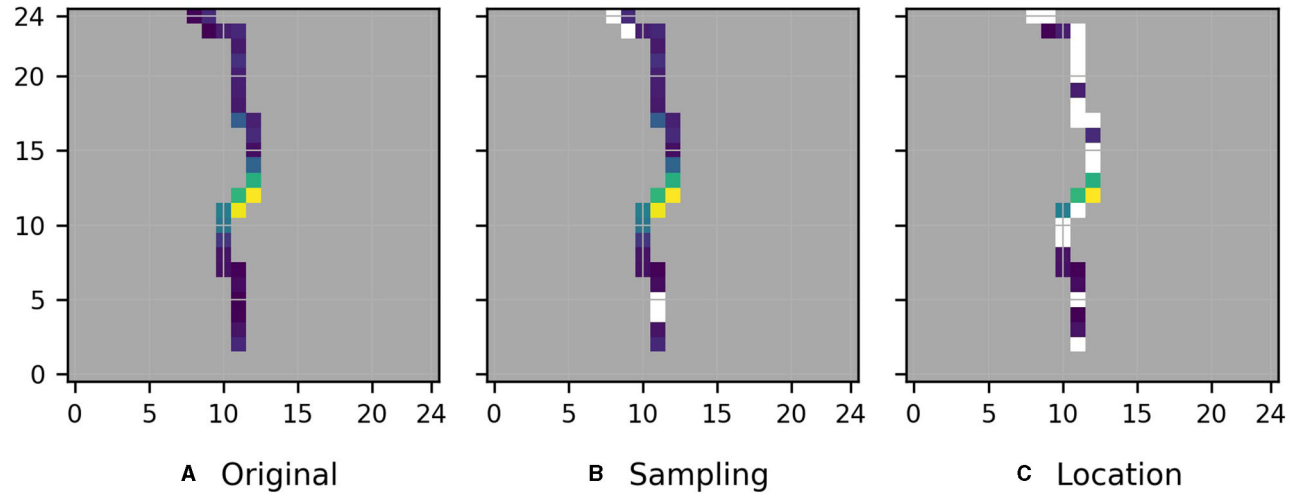

FIGURE 7 | (A) Description $\vec{d}$ of the function that the MLP learned when it was trained on the original water level prediction task (see **Figure 1B**). The description is a measure of the average importance of each pixel in the considered river segment for the predictions of the model. Yellow color indicates high and blue color low importance. **(B,C)** As **(A)**, but pixels for which the relative distance between the original description $\vec{d}$ and one of the sampling and (shifted) location variant descriptions $\vec{d}^{v_i}$, respectively, exceeds the threshold of $t = 0.5$, are masked. Gray marks pixels outside the considered river segment.

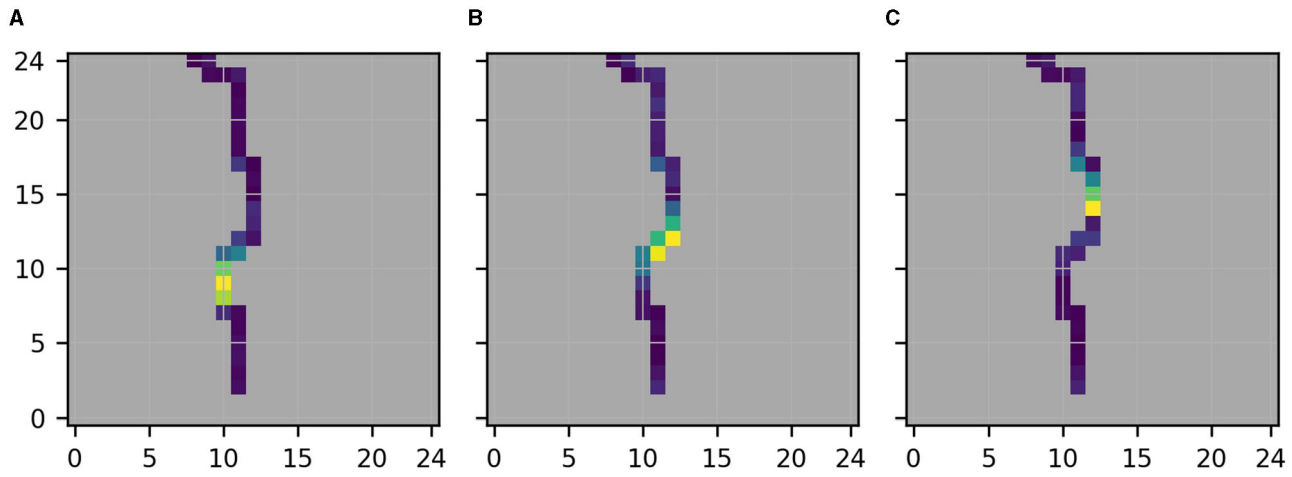

FIGURE 8 | Descriptions of the functions that separate instances of the MLP learned when trained for the different target locations [From left to right the target location is at (10,10), (12,12), (12,15), see **Figure 2B**]. Note that **(B)** shows the same as **Figure 7A**. Gray marks pixels outside the considered river segment. **(A)** Lower target location. **(B)** Central target location. **(C)** Upper target location.

agrees well with the visual assessment of the location variant descriptions depicted in **Figure 10**. Indeed, visual assessment of **Figure 10** also indicates that the neighboring pixel upstream of the target location and maybe the target location itself are the only two pixels for which the assigned importance approximately reflects the shift of the target location between **Figures 10A–C**.

4. CONCLUSIONS

Given a description $\vec{d} \in \mathbb{R}^d$ of the function that a statistical model learned during a training phase, we proposed a variant approach for the identification of parts of $\vec{d}$ that reflect spurious relations. We successfully demonstrated the approach and its superiority over pure sampling approaches with two illustrative hydrometeorological predictions tasks, various statistical models and illustrative descriptions. For the rain prediction task, where we assumed causal relations to remain stable between original and variant tasks, the formal evaluation of distances between original and variant descriptions enabled us to correctly identify all spurious relations that the statistical models learned. For the water level prediction task, where formally specifying the assumed variation of causal relations was more involved, we

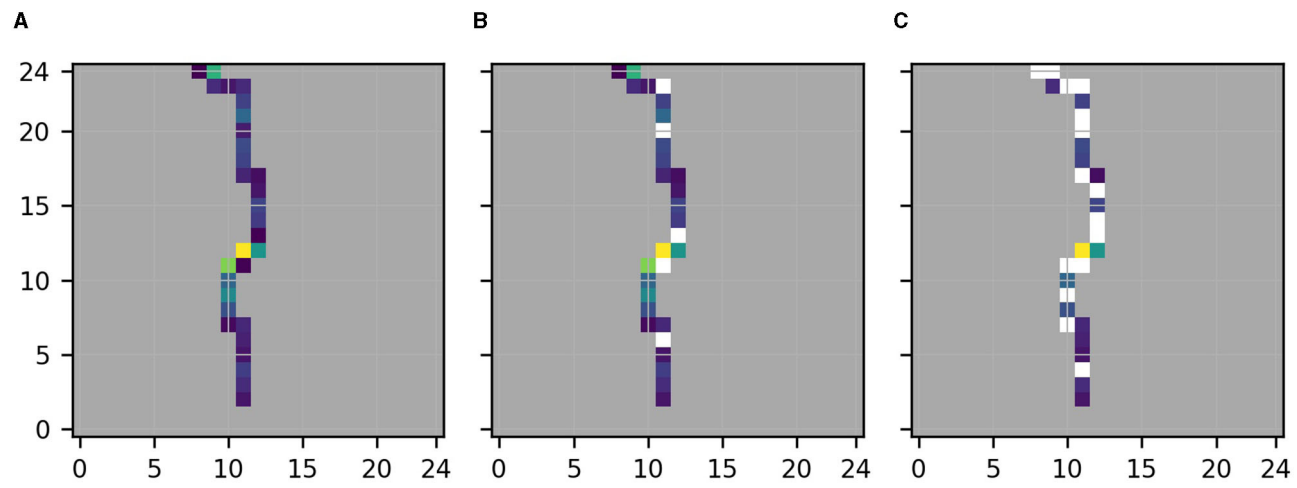

FIGURE 9 | Same as **Figure 7** but for the linear regression model. **(A)** Original. **(B)** Sampling. **(C)** Location.

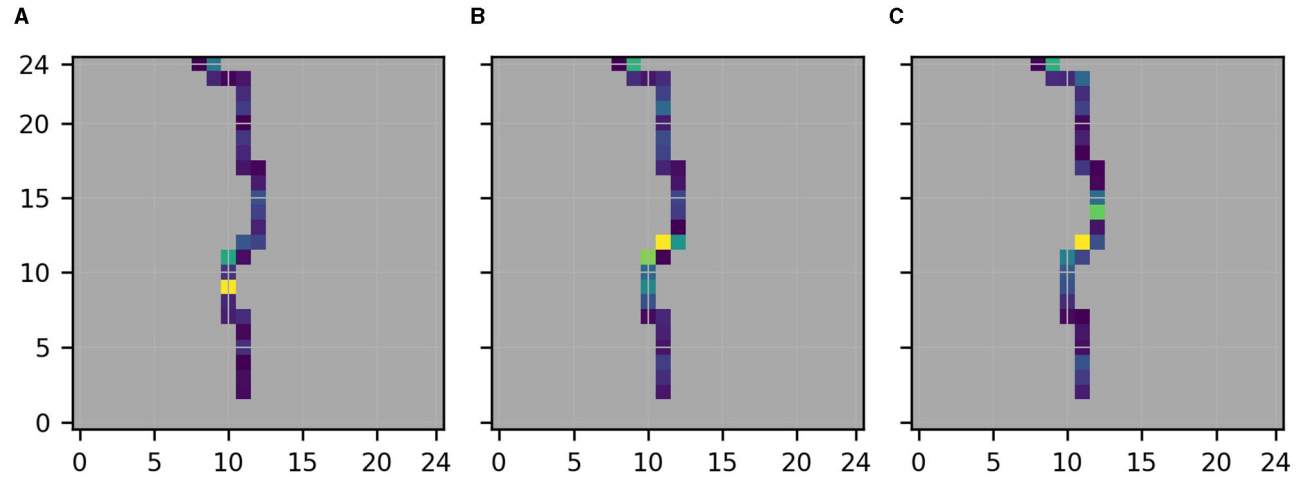

FIGURE 10 | Same as **Figure 8** but for the linear regression model. **(A)** Lower target location. **(B)** Central target location. **(C)** Upper target location.

found the formal evaluation of distances to be of limited use. However, visual assessment enabled us again to correctly identify all spurious relations that the statistical models learned.

In this work, we considered simplified tasks and global descriptions of the learned functions to be able to decide whether parts of the descriptions that the variant approach identifies as spurious do indeed reflect spurious relations. This was necessary to evaluate the variant approach. However, we expect the approach to be beneficial for a wide range of more complex prediction tasks. Naming two possible applications outside the geosciences, it might be used to identify spurious relations reflected in (local) descriptions of functions that DL models trained on electroencephalography (EEG) data (Sturm et al., 2016) learned by comparing them to variant descriptions obtained for variant models trained for different (groups of) patients; or to automatically detect spurious relations reflected in (local) descriptions of functions that a DL model trained on a common image data set learned (Lapuschkin et al., 2019) by automatically comparing them to variant descriptions for variant models trained on different image data sets. Applications of the variant approach to more complex prediction tasks in the geosciences and beyond, and to local descriptions of the learned functions, are planned in future.

A challenge when applying the proposed variant approach may be to define variant tasks beyond random sampling of the training data. However, a data set is often composed of different sources constituting in themselves variants. Further, the modification of the rain prediction task, where we were able to identify parts of the original description as spurious by merely changing the size of the input region, indicates that even small modifications of the original prediction task can be useful.

Apart from the variant approach, which considers a fixed statistical model and modifications of an original prediction task, another approach for identifying spurious relations that a considered statistical model learned might be to compare the relations between input and target variables that different models learn when trained on the (fixed) prediction task. In such an approach, the degree of variation between models may differ from varying configurations in Monte-Carlo dropout, over random seeds for the weight initialization of otherwise identical models to completely different statistical models. Formalization and evaluation of this approach is out of scope of this work.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. Code for replication of the study is available at: https://datapub.fz-juelich.de/slts/t_tesch/.

AUTHOR CONTRIBUTIONS

TT and SK designed the experiments. TT conducted the experiments, analyzed the results and prepared the manuscript with contributions from SK and JG. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors 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 center (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.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frwa.2021.745563/full#supplementary-material

REFERENCES

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* 10:e0130140. doi: 10.1371/journal.pone.0130140

De Bin, R., Janitza, S., Sauerbrei, W., and Boulesteix, A.-L. (2015). Subsampling versus bootstrapping in resampling-based model selection for multivariable regression. *Biometrics* 72, 272–280. doi: 10.1111/biom.12381

Furusho-Percot, C., Goergen, K., Hartick, C., Kulkarni, K., Keune, J., and Kollet, S. (2019). Pan-european groundwater to atmosphere terrestrial systems climatology from a physically consistent simulation. *Sci. Data* 6:320. doi: 10.1038/s41597-019-0328-7

Gagne, D. J. II, Haupt, S. E., Nychka, D. W., and Thompson, G. (2019). Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Rev*. 147, 2827–2845. doi: 10.1175/MWR-D-18-0316.1

Gasper, F., Goergen, K., Shrestha, P., Sulis, M., Rihani, J., Geimer, M., et al. (2014). Implementation and scaling of the fully coupled terrestrial systems modeling platform (TerrSysMP v1.0) in a massively parallel supercomputing environment– a case study on JUQUEEN (IBM Blue Gene/Q). *Geosci. Model Dev*. 7, 2531–2543. doi: 10.5194/gmd-7-2531-2014

Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M., and Kagal, L. (2018). "Explaining explanations: an overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)* (Turin: IEEE), 80–89. doi: 10.1109/DSAA.2018.00018

Goodfellow, I., Shlens, J., and Szegedy, C. (2015). "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*. San Diego, CA.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. (2020). A survey of learning causality with data: problems and methods. *ACM Comput. Surv*. 75, 1–37. doi: 10.1145/3397269

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

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 770–778. doi: 10.1109/CVPR.2016.90

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning* (Lille: PMLR), 448–456.

Lakkaraju, H., Arsov, N., and Bastani, O. (2020). "Robust and stable black box explanations," in *Proceedings of the 37th International Conference on Machine Learning* (PMLR) 119, 5628–5638.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nat. Commun*. 10:1096. doi: 10.1038/s41467-019-08987-4

Larraondo, P. R., Renzullo, L. J., Inza, I., and Lozano, J. A. (2019). A data-driven approach to precipitation parameterizations using convolutional encoder-decoder neural networks. *arXiv*.

LeCun, Y., Bottou, L., Orr, G., and Müller, K. (2012). *Efficient BackProp*. Berlin; Heidelberg: Springer. doi: 10.1007/978-3-642-35289-8_3

McGovern, A., Lagerquist, R., Gagne, I. I., D. J., Jergensen, G. E., Elmore, K. L., et al. (2019). Making the black box more transparent: understanding the physical implications of machine learning. *Bull. Am. Meteor. Soc*. 100, 2175–2199. doi: 10.1175/BAMS-D-18-0195.1

Molnar, C. (2019). Interpretable machine learning. *A Guide for Making Black Box Models Explainable*. doi: 10.21105/joss.00786. Available online at: https://christophm.github.io/interpretable-ml-book/

Montavon, G., Samek, W., and Müller, K. (2018). Methods for interpreting and understanding deep neural networks. *Digit. Signal Process*. 73, 1–15. doi: 10.1016/j.dsp.2017.10.011

Odena, A., Dumoulin, V., and Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*. doi: 10.23915/distill.00003

Pan, B., Hsu, K., AghaKouchak, A., and Sorooshian, S. (2019). Improving precipitation estimation using convolutional neural network. *Water Resour. Res*. 55, 2301–2321. doi: 10.1029/2018WR024090

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng*. 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "Pytorch: an imperative style, high-performance deep learning library," in

Advances in Neural Information Processing Systems 32, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (New York, NY: Curran Associates, Inc.), 8026–8037.

Pearl, J. (2009). Causal inference in statistics: an overview. *Stat. Surv.* 3, 96–146. doi: 10.1214/09-SS057

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830.

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

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ""Why should I trust you?": explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16* (New York, NY: Association for Computing Machinery), 1135–1144. doi: 10.1145/2939672.2939778

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access* 8, 42200–42216. doi: 10.1109/ACCESS.2020.2976199

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., and Müller, K. R. (2021). Explaining deep neural networks and beyond: a review of methods and applications. *Proc. IEEE* 109, 247–278. doi: 10.1109/JPROC.2021.3060483

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., et al. (2020). Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.* 2, 476–486. doi: 10.1038/s42256-020-0212-3

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice: IEEE), 618–626. doi: 10.1109/ICCV.2017.74

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference* 90, 227–244. doi: 10.1016/S0378-3758(00)00115-4

Shrestha, P., Sulis, M., Masbou, M., Kollet, S., and Simmer, C. (2014). A scale-consistent terrestrial systems modeling platform based on COSMO, CLM, and ParFlow. *Monthly Weather Rev.* 142, 3466–3483. doi: 10.1175/MWR-D-14-00029.1

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). "Deep inside convolutional networks: visualising image classification models and saliency maps," in *Workshop at International Conference on Learning Representations* (Banff). Available online at: https://arxiv.org/abs/1312.6034

Sinha, A., Namkoong, H., and Duchi, J. C. (2018). "Certifiable distributional robustness with principled adversarial training," in *International Conference on Learning Representations*. Vancouver.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Sturm, I., Lapuschkin, S., Samek, W., and Müller, K. (2016). Interpretable deep neural networks for single-trial EEG classification. *J. Neurosci. Methods* 274, 141–145. doi: 10.1016/j.jneumeth.2016.10.008

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2014). "Intriguing properties of neural networks," in *International Conference on Learning Representations*. Venice.

Toms, B. A., Barnes, E. A., and Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: applications to earth system variability. *J. Adv. Model. Earth Syst.* 12:e2019MS002002. doi: 10.1029/2019MS002002

Zhang, Q., and Zhu, S. (2018). Visual interpretability for deep learning: a survey. *Front. Inf. Technol. Electron. Eng.* 19, 27–39. doi: 10.1631/FITEE.1700808

Author Disclaimer: 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.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

A.2. Supporting information

Supplementary Material

1 MODEL ARCHITECTURES AND TRAINING

1.1 Multilayer Perceptron (MLP)

The MLP considered for the water level prediction task receives as input 25×25 pixels water level fields flattened to vectors $\vec{x} \in \mathbb{R}^{25 \cdot 25}$. Those vectors are passed through a hidden layer with 100 neurons and a rectified linear unit (ReLU) nonlinear activation function to an output layer with one neuron whose linear activation is the prediction of the network. The target $y \in \mathbb{R}$ is the water level at the target location normalized to have a mean of 0 and a standard deviation of 1 on the training set.

We trained the network for 150 epochs using the Adam optimizer with the learning rate and weight decay parameters both set to 1e-3, a batch size of 256 and the mean squared error as loss function. We implemented this using the deep learning library Pytorch (Paszke et al., 2019).

1.2 Convolutional Neural Networks (CNNs)

The input to the CNNs are geopotential fields at 500, 850 and 1000 hPa. For each pressure level, we normalized the geopotential fields to have a mean of 0 and a standard deviation of 1 on the training set.

CNN1 is a shallow CNN with two convolutional layers, each consisting of 24 kernels of size 3 and rectified linear units (ReLU) nonlinear activation functions. The convolutional layers are followed by a fully-connected layer connecting the second convolutional layer to two output neurons. CNN2 is a commonly employed, much deeper CNN architecture called resnet18 (He et al., 2016) for which the last fully-connected layer was adapted to have only two output neurons to fit our prediction task.

The networks predict a rain event for a given input if the value of the second output neuron is higher than the value of the first output neuron. To obtain a 2-dimensional vector with no rain and rain probabilities which sum to one, a softmax function is applied

$$P(\text{rain event})_i = \frac{\exp(out_i)}{\exp(out_1) + \exp(out_2)}, \tag{S1}$$

where out_i is the activation of the i-th output neuron. Note that for the saliency maps, we consider the derivative of the output of the second output neuron before the application of the softmax layer. Applying the saliency maps to the first output neuron yields very similar results. The target vector for an input $\vec{x}$ is a 2-dimensional vector indicating whether a rain event occurred ($\vec{y} = (0, 1)$) or not ($\vec{y} = (1, 0)$).

We trained both networks for 60 epochs using the Adam optimizer with a learning rate of 1e-3, a batch size of 1000, the CrossEntropyLoss-criterion and the ReduceLROnPlateau scheduler with patience parameter set to 6. We implemented this using the deep learning library Pytorch (Paszke et al., 2019).

Note that for CNN1 the number of neurons in the fully-connected layer has to be adapted when the size of the input region is changed to 80×80 pixels. This is not the case for the resnet18 model, as for this model, the output of the last convolutional layer is always pooled to a fixed size.

REFERENCES

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. doi:10.1109/CVPR.2016.90

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, eds. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc.). 8026–8037

B. Causal deep learning models for studying the Earth system

Originally published in

T. Tesch, S. Kollet, and J. Garcke. Causal deep learning models for studying the Earth system. *Geoscientific Model Development*, 16(8):2149–2166, 2023a. doi: 10.5194/gmd-16-2149-2023. Highlight article.

Geosci. Model Dev., 16, 2149–2166, 2023
https://doi.org/10.5194/gmd-16-2149-2023

Geoscientific Model Development
Open Access
EGU

Causal deep learning models for studying the Earth system

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

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

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

Received: 26 March 2022 – Discussion started: 17 May 2022
Revised: 30 November 2022 – Accepted: 8 February 2023 – Published: 20 April 2023

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

1 Introduction

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

Approaches for studying relations in the Earth system may be broadly divided into approaches based on numerical simulations (e.g., Koster, 2004; Seneviratne et al., 2006; Hartick et al., 2021) and statistical approaches (e.g., Taylor, 2015; Guillod et al., 2015; Tuttle and Salvucci, 2016). Both classes of approaches have several inherent limitations. Approaches based on numerical simulations usually have high computational costs and, even more importantly, rely on the correct representation of the considered relations in the numerical model. For example, precipitation in numerical models lacks accuracy due to several simplified parameterizations; thus, using these models to study soil-moisture–precipitation coupling is problematic. On the other hand, statistical approaches usually have much lower computational costs and can directly be applied to observational data. However, current statistical approaches have strong limitations on their

own, for example, due to assumptions on linearity or locality of considered relations and negligence of the difference between causality and correlation.

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

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

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

2 Methodology

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

2.1 Background on causality

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

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

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

2.1.1 Structural causal models

In the framework of SCMs, the considered system, e.g., the Earth system, is described by a causal graph and associated structural equations. A causal graph is a directed acyclic graph, in which nodes represent the variables of the system and edges encode the dependencies between these variables. For example, in the system described by Fig. 1a, variable $\boldsymbol{Y}$ depends on all other variables, although the lack of an edge from $\boldsymbol{X}$ to $\boldsymbol{Y}$ implies that $\boldsymbol{X}$ only affects $\boldsymbol{Y}$ indirectly via its impact on $\boldsymbol{C}_2$. Parents of a considered variable (node) are all variables that have a direct effect on that variable, i.e., all variables with an edge pointing to that variable. In the follow-

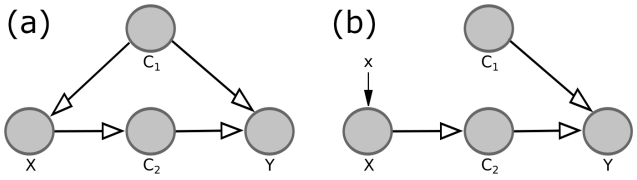

Figure 1. Example for a causal graph **(a)** and corresponding causal graph for setting variable $\boldsymbol{X}$ to some arbitrary value $\boldsymbol{x}$ **(b)**. The grey circles are referred to as nodes of the graph, while the arrows are referred to as directed edges.

ing, the terms "node" and "variable" are used interchangeably.

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

$$\begin{aligned} &\boldsymbol{C}_1 = f_{C_1}(\boldsymbol{U}_{C_1}),\ \boldsymbol{X} = f_X(\boldsymbol{C}_1, \boldsymbol{U}_X), \\ &\boldsymbol{C}_2 = f_{C_2}(\boldsymbol{X}, \boldsymbol{U}_{C_2}),\ \boldsymbol{Y} = f_Y(\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{U}_Y). \end{aligned} \tag{1}$$

These equations are called structural equations. The random variables $\boldsymbol{U}_{C_1}, \boldsymbol{U}_{C_2}, \boldsymbol{U}_X, \boldsymbol{U}_Y$ are assumed to be mutually independent and give rise to a probability distribution $\mathbb{P}(\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{X}, \boldsymbol{Y})$, which describes the probability of observing any tuple of values $(\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{x}, \boldsymbol{y})$. Integrating the product of $\boldsymbol{Y}$ and this probability distribution over all tuples $(\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{y})$ for some fixed value $\boldsymbol{x}$, one obtains the expected value of $\boldsymbol{Y}$ given that one observes the value $\boldsymbol{x}$ of $\boldsymbol{X}$, i.e.,

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}] = \int_{\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{y}} \boldsymbol{y} \cdot \mathbb{P}[\boldsymbol{C}_1 = \boldsymbol{c}_1, \boldsymbol{C}_2 = \boldsymbol{c}_2, \boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X} = \boldsymbol{x}]. \tag{2}$$

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

$$\begin{aligned} &\boldsymbol{C}_1 = f_{C_1}(\boldsymbol{U}_{C_1}),\ \boldsymbol{X} = \boldsymbol{x},\ \boldsymbol{C}_2 = f_{C_2}(\boldsymbol{X}, \boldsymbol{U}_{C_2}), \\ &\boldsymbol{Y} = f_Y(\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{U}_Y). \end{aligned} \tag{3}$$

Again, the random variables $\boldsymbol{U}_{C_1}, \boldsymbol{U}_{C_2}, \boldsymbol{U}_Y$ give rise to a probability distribution $\mathbb{P}(\boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{Y}|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x}))$, referred to as the post-intervention probability distribution, and the corresponding post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x})]$. This expected value is used to determine the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ and differs from the expected value for the original system, $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}]$. For instance, in the example from Fig. 1, knowing $\boldsymbol{X}$ allows us to draw conclusions about $\boldsymbol{Y}$ both in the original system (Fig. 1a) and in the modified system (Fig. 1b), because $\boldsymbol{X}$ has a causal effect on $\boldsymbol{Y}$ (via its impact on $\boldsymbol{C}_2$). However, in the original system, knowing $\boldsymbol{X}$ allows us to draw additional conclusions about $\boldsymbol{C}_1$. This is the case although the edge in the causal graph points from $\boldsymbol{C}_1$ to $\boldsymbol{X}$; i.e., $\boldsymbol{C}_1$ affects $\boldsymbol{X}$, not vice versa. For example, if $\boldsymbol{X}$ was simply the sum of $\boldsymbol{C}_1$ and the random term $\boldsymbol{U}_X$, a high value of $\boldsymbol{X}$ would probably imply a high value of $\boldsymbol{C}_1$. These conclusions about $\boldsymbol{C}_1$ cannot be drawn in the modified system, where the edge from $\boldsymbol{C}_1$ to $\boldsymbol{X}$ is removed. The knowledge about $\boldsymbol{C}_1$ allows us to draw further conclusions about $\boldsymbol{Y}$ because $\boldsymbol{C}_1$ also affects $\boldsymbol{Y}$. Summarizing, due to the confounding influence of $\boldsymbol{C}_1$, knowing $\boldsymbol{X}$ reveals more about $\boldsymbol{Y}$ in the original system than in the modified system, which is why the original expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}]$ and the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x})]$ 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 $\mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x})]$ by training a suitable (see Sect. 2.2.1) statistical model on the observed tuples $(\boldsymbol{x}, \boldsymbol{y})$ to predict $\boldsymbol{Y}$ given $\boldsymbol{X}$. However, in the cases considered in the proposed methodology, it is impossible or undesirable to experimentally set $\boldsymbol{X}$ to $\boldsymbol{x}$. Thus, we can only observe the original system and approximate the original expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}]$ by analogously training a statistical model on observed tuples $(\boldsymbol{x}, \boldsymbol{y})$ of the original system. Consequently, we have to bridge the gap between the original expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}]$ and the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x})]$.

2.1.2 Adjustment criteria

To bridge the gap between the original expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}]$ and the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x})]$, we must take into account variables other than $\boldsymbol{X}$ and $\boldsymbol{Y}$. Indeed, in the example from Fig. 1, we showed that original and post-intervention expected values differ because, in the original system, knowing $\boldsymbol{X}$ allows inferences about $\boldsymbol{C}_1$ that are not possible in the modified system. However, if we actually knew $\boldsymbol{C}_1$, this would not be the case; thus, the original expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{C}_1 = \boldsymbol{c}_1]$ and the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{C}_1 = \boldsymbol{c}_1]$ are identical. Analogously to $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}]$, the expected value $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{C}_1 = \boldsymbol{c}_1]$ can be approximated by observing the original system and training a statistical model on the observed tuples $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{c}_1)$ to predict $\boldsymbol{Y}$ given $\boldsymbol{X}$ and $\boldsymbol{C}_1$. Therefore, this equality allows us to approximate the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{C}_1 = \boldsymbol{c}_1]$ by only observing the original system and without experimentally setting $\boldsymbol{X}$ to $\boldsymbol{x}$.

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

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\{\boldsymbol{C}_\ell=\boldsymbol{c}_\ell\}_{\ell=1}^k] = \mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C}_\ell=\boldsymbol{c}_\ell\}_{\ell=1}^k] \quad (4)$$

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

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

Here, a path is any consecutive sequence of edges. A path between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is causal from $\boldsymbol{X}$ to $\boldsymbol{Y}$ if all edges point towards $\boldsymbol{Y}$, and it is non-causal otherwise. A path is blocked by a set $S = \{\boldsymbol{C}_\ell\}_{\ell=1}^k$ of nodes if either (i) the path contains at least one edge-emitting node, i.e., a node with at least one adjacent edge pointing away from the node ($\ldots \leftrightarrow \boldsymbol{C} \rightarrow \ldots$), that is in S (e.g., the path $\boldsymbol{X} \leftarrow \boldsymbol{C}_1 \rightarrow \boldsymbol{Y}$ in Fig. 1 is blocked by S if S contains $\boldsymbol{C}_1$), or (ii) the path contains at least one collision node, i.e., a node with both adjacent edges pointing towards the node ($\ldots \rightarrow \boldsymbol{C} \leftarrow \ldots$), which is outside S and has no descendants in S (e.g., the path $\boldsymbol{X} \rightarrow \boldsymbol{C} \leftarrow \boldsymbol{Y}$ is blocked if S does not contain $\boldsymbol{C}$).

The first adjustment criterion generalizes the example of $\boldsymbol{C}_1$ in Fig. 1, where adjusting for the edge-emitting node $\boldsymbol{C}_1$, i.e., considering $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{C}_1=\boldsymbol{c}_1]$ rather than $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}]$, blocks the non-causal path $\boldsymbol{X} \leftarrow \boldsymbol{C}_1 \rightarrow \boldsymbol{Y}$ such that $\boldsymbol{X}$ is only used to draw conclusions about $\boldsymbol{Y}$ via the causal path $\boldsymbol{X} \rightarrow \boldsymbol{C}_2 \rightarrow \boldsymbol{Y}$. In general, the criterion ensures that $\boldsymbol{X}$ is only used to draw conclusions about $\boldsymbol{Y}$ via causal paths from $\boldsymbol{X}$ to $\boldsymbol{Y}$ and not via any non-causal path between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

The second adjustment criterion ensures that no causal path from $\boldsymbol{X}$ to $\boldsymbol{Y}$ is blocked, such that the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C}_\ell=\boldsymbol{c}_\ell\}_{\ell=1}^k]$ actually reflects the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$. For example, considering the causal path $\boldsymbol{X} \rightarrow \boldsymbol{C}_2 \rightarrow \boldsymbol{Y}$ in Fig. 1, $\boldsymbol{C}_2$ blocks the only causal path between $\boldsymbol{X}$ and $\boldsymbol{Y}$. Thus, $\mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X}=\boldsymbol{x}),\boldsymbol{C}_2=\boldsymbol{c}_2] = \mathbb{E}[\boldsymbol{Y}|\boldsymbol{C}_2=\boldsymbol{c}_2]$ 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 $\mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C}_\ell=\boldsymbol{c}_\ell\}_{\ell=1}^k]$ 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 fulfill the above adjustment criteria. Describing the system by a causal graph requires knowledge on which variables are relevant to the considered relation (represented by the nodes in the graph) and on the existence of causal dependencies between these variables (represented by the edges in the graph). Nevertheless, it does not require knowledge on the sign or strength of these dependencies, i.e., on the structural equations. Note that the parents of $\boldsymbol{X}$ in the causal graph always fulfill the adjustment criteria. In the proposed methodology, we exploit the post-intervention expected value $\mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C}_\ell=\boldsymbol{c}_\ell\}_{\ell=1}^k]$ to determine the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ as detailed in Sect. 2.2.2.

2.2 Steps of the methodology

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

2.2.1 Training a causal DL model

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

$$(\boldsymbol{x},\{\boldsymbol{c}_\ell\}_{\ell=1}^k) \rightarrow \mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X}=\boldsymbol{x}),\{\boldsymbol{C}_\ell=\boldsymbol{c}_\ell\}_{\ell=1}^k]. \quad (5)$$

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

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

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

$$(\boldsymbol{m}:(\boldsymbol{X},\{\boldsymbol{C}_\ell\}_{\ell=1}^k) \rightarrow \mathbb{R}^n) \rightarrow \mathbb{E}[(\boldsymbol{Y}-\boldsymbol{m}(\boldsymbol{x},\{\boldsymbol{c}_\ell\}_{\ell=1}^k))^2], \quad (7)$$

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 $\{\boldsymbol{C}_\ell\}_{\ell=1}^k$ that fulfill the adjustment criteria from Sect. 2.1.2, such that the maps from Eqs. (5) and (6) become identical. The choice of additional input variables requires prior knowledge on which variables are relevant for the considered relation and on the existence of causal dependencies between these variables. However, it does not require prior knowledge on the strength, sign, or functional form of these dependencies (see Sect. 2.1.2), which can be obtained from the proposed methodology.

2.2.2 Sensitivity analysis of the trained model

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

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

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

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

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

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

In theory, the proposed methodology identifies the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ exactly. In practice, however, we make two important approximations. First, due to the complexity of the Earth system, the additional input variables $\{\boldsymbol{C}_\ell\}_{\ell=1}^k$ may not strictly fulfill the adjustment criteria from Sect. 2.1.2, such that the map from Eq. (6) is only approximately identical to the map from Eq. (5). Second, the DL model only approximates the map from Eq. (6). Thus, the partial derivatives $q_{ij}(\boldsymbol{x}, \{\boldsymbol{c}_\ell\}_{\ell=1}^k)$ of the DL model only approximate the partial derivatives $s_{ij}(\boldsymbol{x}, \{\boldsymbol{c}_\ell\}_{\ell=1}^k)$ that we are interested in. We will come back to this in Sects. 3.3 and 4.

3 Application to soil-moisture–precipitation coupling

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

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

3.1 Data

The data underlying our example are ERA5 hourly data (Hersbach et al., 2023) constituting an atmospheric reanalysis of the past decades (1950 to today) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). Reanalysis means simulation data and observations have been merged into a single description of the global climate and weather using data assimilation technologies. ERA5 data contain hourly estimates for a large number of atmospheric, ocean-wave and land-surface quantities on a regular latitude–longitude grid of 0.25° ($\approx 30\,\mathrm{km}$). In this study, soil moisture refers to the ERA5 variable "volumetric soil water in the upper soil layer (0–7 cm)". The target variable, precipitation $\boldsymbol{P}[t+4\,\mathrm{h}]$, represents an accumulation of precipitation over the time interval $[t + 3\,\mathrm{h}, t + 4\,\mathrm{h}]$. 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, $\boldsymbol{P}[t + 4\,\mathrm{h}]$, for times $t + 4\,\mathrm{h}$ between noon and 23:00 UTC.

3.2 Loss function, model architecture and training

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

$$(\boldsymbol{SM}[t], \{\boldsymbol{C}_\ell[t]\}_{\ell=1}^k) \to \mathbb{E}[\boldsymbol{P}[t + 4\,\mathrm{h}] | \boldsymbol{SM}[t], \{\boldsymbol{C}_\ell[t]\}_{\ell=1}^k]. \quad (10)$$

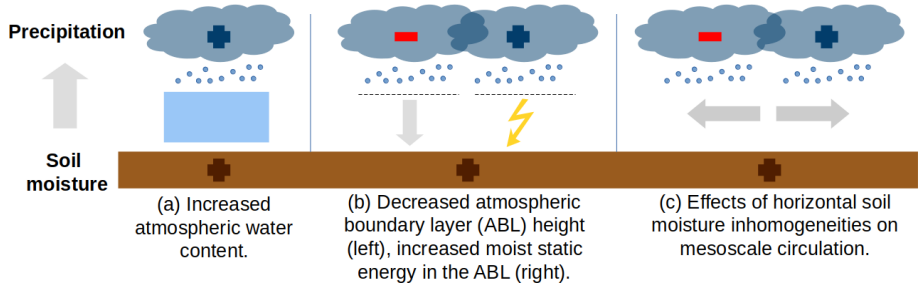

Figure 2. Concurring pathways of soil-moisture–precipitation coupling. An increase in soil moisture can increase latent heat flux and decrease sensible heat flux at the land surface (Seneviratne et al., 2010). This can increase precipitation via an increase in atmospheric water content (**a**; Eltahir,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

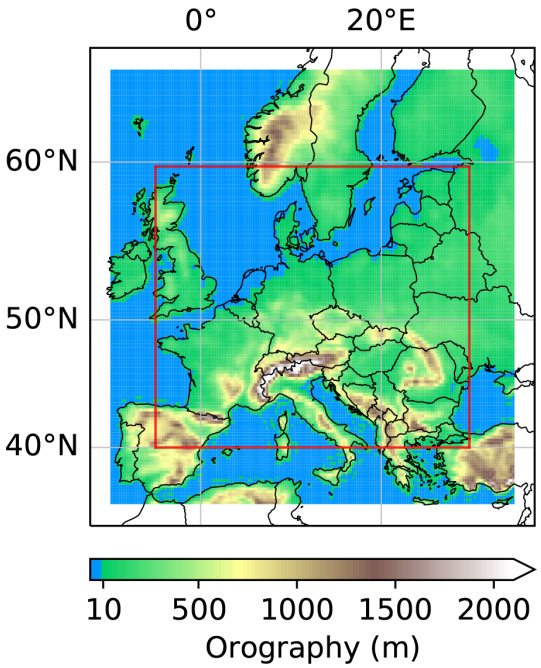

Figure 3. Input and target regions in the example of soil-moisture–precipitation coupling. The colored region represents the 120×180 pixel input region and the red box the 80×140 pixel target region. Note that the offset between input and target region is 20 pixels on each side and distorted by the projection.

This holds true for the expected mean squared error from Eq. (7). Given N training time steps t_i, associated values $(\boldsymbol{SM}[t_i], \{\boldsymbol{C}_\ell[t_i]\}_{\ell=1}^k, \boldsymbol{P}[t_i + 4\,\mathrm{h}])_{i=1}^N$ 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

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

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

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

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

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

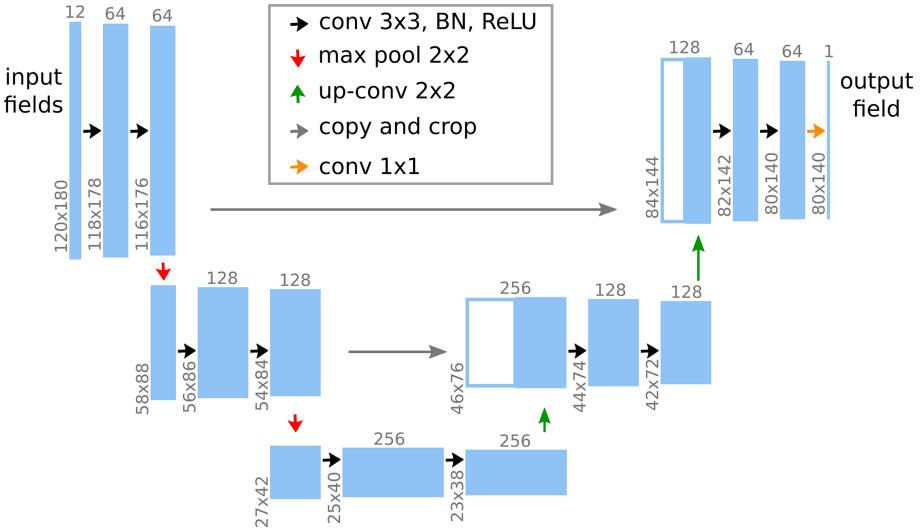

Figure 4. Model architecture in the example of soil-moisture–precipitation coupling. The leftmost blue box represents the input to the model, which consists of 12 variables (including soil moisture) at the 120×180 input pixels (see Fig. 3). This input is passed through multiple sequential modules represented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

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

During training, the Adam optimizer (Kingma and Ba, 2017) is used to adapt the approximately 2.3 million, randomly initialized weights of the model to minimize the mean squared error on the training set. In terms of implementation, we use the PyTorch (Paszke et al., 2019) wrapper skorch (Tietz et al., 2017) with default parameters for training the model: set the maximum number of epochs to 200, the learning rate in the Adam optimizer to 1×10^{-3}, the batch size to 64, and patience for early stopping (i.e., the number of epochs after which training stops if the loss function evaluated on the validation set does not improve by some threshold) to 30 epochs. During training, we further use data augmentation. Namely, we randomly rotate by 180° (or not) and subsequently horizontally flip (or not) the considered region for each training sample and each training epoch independently. Similar to the translation invariance of the model, this requires including input variables which lead to spatial variability in soil-moisture–precipitation coupling as discussed in the next section.

3.3 Choice of input variables

The choice of additional input variables $\{\boldsymbol{C}_\ell\}_{\ell=1}^k$ represents a crucial aspect of the proposed methodology for two reasons (see Sect. 2.2.2). First, we need the additional input variables to (approximately) fulfill 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 $\mathbb{E}[\boldsymbol{P}[t+4\,\mathrm{h}]|\boldsymbol{SM}[t], \{\boldsymbol{C}_\ell[t]\}_{\ell=1}^k]$ (see Eq. 10) is a good approximation of the map

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

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 $\boldsymbol{SM}[t]$ values that are computed in the sensitivity analysis (see Sect. 3.4).

Choosing additional input variables that fulfill 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 fulfill 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 timescale of our analysis, a reasonable example is the set of variables in the second column in Fig. 5.

If we included all of these variables, the adjustment criteria would be met and the map from Eq. (10) would be

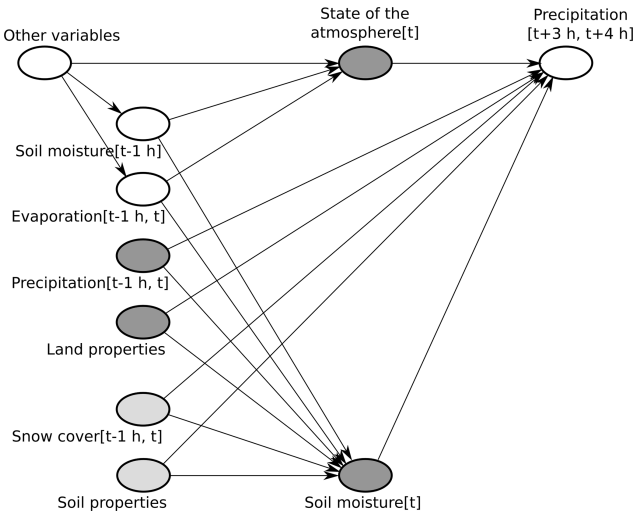

Figure 5. Causal graph in the example of soil-moisture–precipitation coupling. The dark grey nodes represent the chosen input variables, while light grey nodes represent variables that are ignored in our analysis (see text). Land properties comprise the time-independent variables topography, land–sea 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 $[t, t+3\,\mathrm{h}]$).

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

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

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

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

3.4 Sensitivity analysis

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

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

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

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

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

Note that most of the derivatives in the sum are zero, because, e.g., a change in soil moisture in Great Britain at time t does not affect precipitation in Italy 4 h later. Outside of a 52×52 pixel surrounding region, this is enforced by the architecture of the DL model (see Sect. 3.2), and inside of this region, it is learned during training of the model. As for local soil-moisture–precipitation coupling, $\overline{q^{\mathrm{reg}}}_{ij}$ denotes the average of q_{ij}^{reg} over all input samples from the test set.

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

3.5 Comparison to other approaches

A common approach for studying relations in the Earth system is to consider the linear correlation between variables (Froidevaux et al., 2014; Welty and Zeng, 2018; Holgate et al., 2019). Here, we compare our results on regional soil-moisture–precipitation coupling to results obtained from a linear correlation analysis. For each location in the considered target region, Fig. 7 shows the linear correlation coefficient of soil moisture $\boldsymbol{SM}[t]$ at that location and subsequent precipitation $\boldsymbol{P}[t + 4\,\mathrm{h}]$ summed over the 15×15 pixel 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 non-linear effects and for the difference between causality and correlation in the proposed methodology.

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

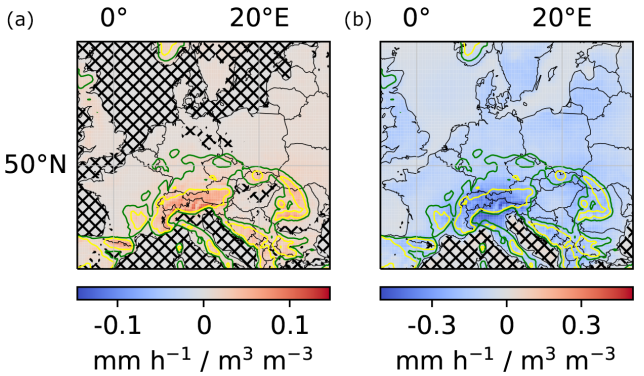

Figure 6. Local and regional soil-moisture–precipitation couplings. **(a)** Impact of local soil moisture changes (m^3 water m^{-3} soil) on local precipitation ($\mathrm{mm\,h}^{-1}$) for each pixel in the target region (in the text denoted by $\overline{q^{\mathrm{loc}}}_{ij}$). **(b)** Impact of local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

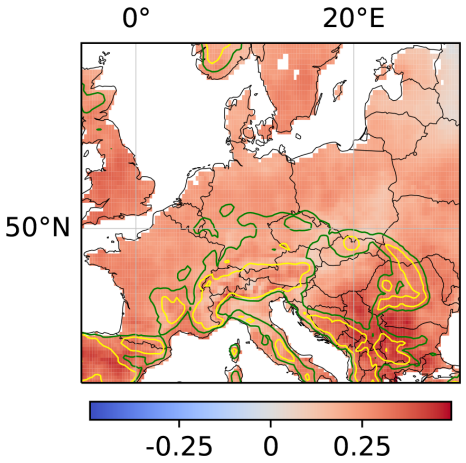

Figure 7. Linear correlation coefficient of local soil moisture and regional precipitation. For each location, the linear correlation coefficient of soil moisture $\boldsymbol{SM}[t]$ at the location and subsequent precipitation $\boldsymbol{P}[t + 4\,\mathrm{h}]$ summed over the 15×15 pixel surrounding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

of 80×140 target pixels on subsequent precipitation in the target region. We averaged these effects over all time steps in 2 test years, constituting 2208 time steps. Performing an analogous study based on numerical simulations would require at least $80 \times 140 \times 2208 = 24\,729\,600$ 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 2208 considered time steps, with the only difference being that soil moisture would be slightly increased or decreased at one of the 80×140 target pixels. This corresponds to simulating more than 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

To ensure that the proposed methodology provides reliable results, this section presents some additional analyses. Theoretically, the proposed methodology determines the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ exactly. However, in practice, we make two important approximations (see Sect. 2.2.2). First, the additional input variables $\{\boldsymbol{C}_\ell\}_{\ell=1}^k$ may not strictly fulfill 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 $\mathbb{E}[\boldsymbol{Y}|\mathrm{do}(\boldsymbol{X} = \boldsymbol{x}), \{\boldsymbol{C}_\ell = \boldsymbol{c}_\ell\}_{\ell=1}^k]$ in Eq. (5). Second, the DL model represents only an approximation of the map from Eq. (6). Both errors are difficult to quantify, because both maps are unknown.

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

The results of the proposed methodology are the partial derivatives $\overline{q_{ij}}$ of the DL model computed in the sensitivity analysis. Due to the above approximations, these derivatives are only approximations of the partial derivatives $\overline{s_{ij}}$ of the map from Eq. (5), which represent the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ (see Sect. 2.2.2). However, even quantifying the two approximation errors mentioned above would not give us a good estimate of the errors in these results. In this section, we propose several additional analyses to build confidence in results obtained with the proposed methodology. Particularly, the proposed analyses show if results are statistically significant, i.e., reflect more than random correlations or artifacts of the DL training procedure (Sect. 4.1), and if they reflect more than specific (known) correlations (Sect. 4.2). Moreover, the analyses proposed in Sect. 4.3 allow us to identify (potentially unknown) spurious correlations in the results. Finally, we propose some further sanity checks in Sect. 4.4. We illustrate the analyses with our results on soil-moisture–precipitation coupling from Sect. 3.

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

4.1 Statistical significance

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

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

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

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

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

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

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

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

4.2 Known spurious correlations

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

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

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

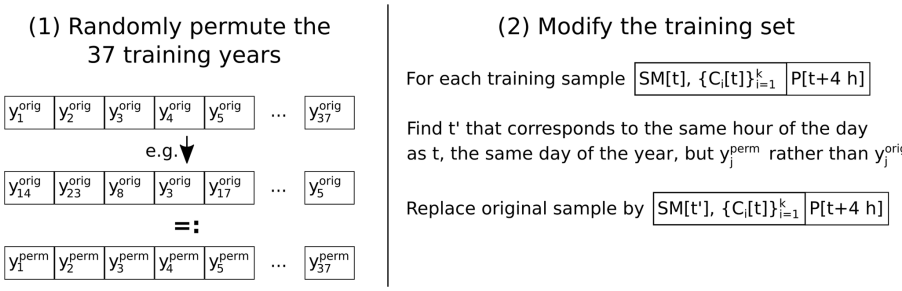

Figure 8. Modification of the training data for the year-wise permutation of $\boldsymbol{SM}[t]$. The modification of the test data works analogously.

lations. Analogously, if the second analysis indicates that the DL model learned useful information in terms of predictive performance on the relation between $\boldsymbol{X}$ and $\boldsymbol{Y}$, apart from the considered spurious correlations, it cannot exclude that the predictions are partly affected by the considered spurious correlations.

4.3 Further spurious correlations

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

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

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

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

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

4.4 Task-specific sanity checks

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

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

4.5 Control experiment

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

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

5 Conclusions

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

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

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

Appendix A

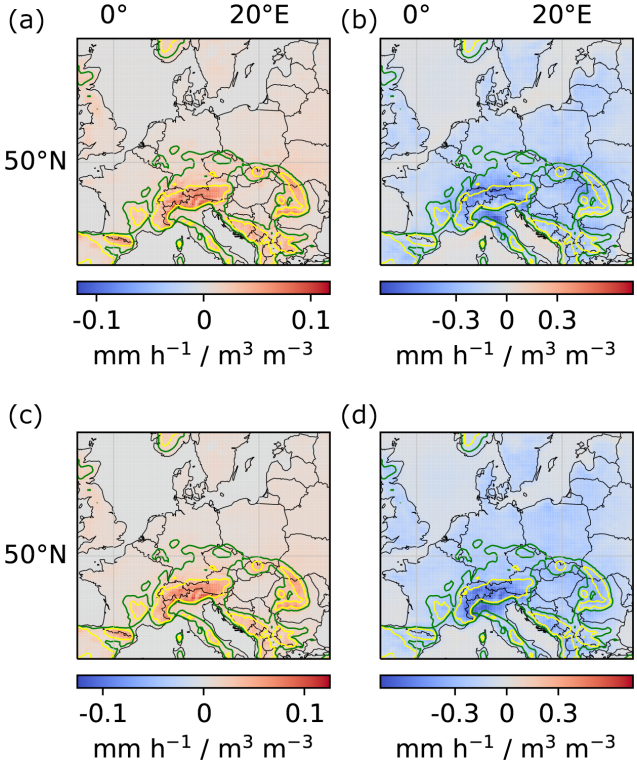

Figure A1. Local and regional soil-moisture–precipitation couplings for models trained on the first and second half of the training years, respectively. **(a, c)** Local couplings. **(b, d)** Regional couplings. **(a, b)** Model trained on the first half of all training years (1979–1997). **(c, d)** Model trained on the second half of all training years (1998–2019).

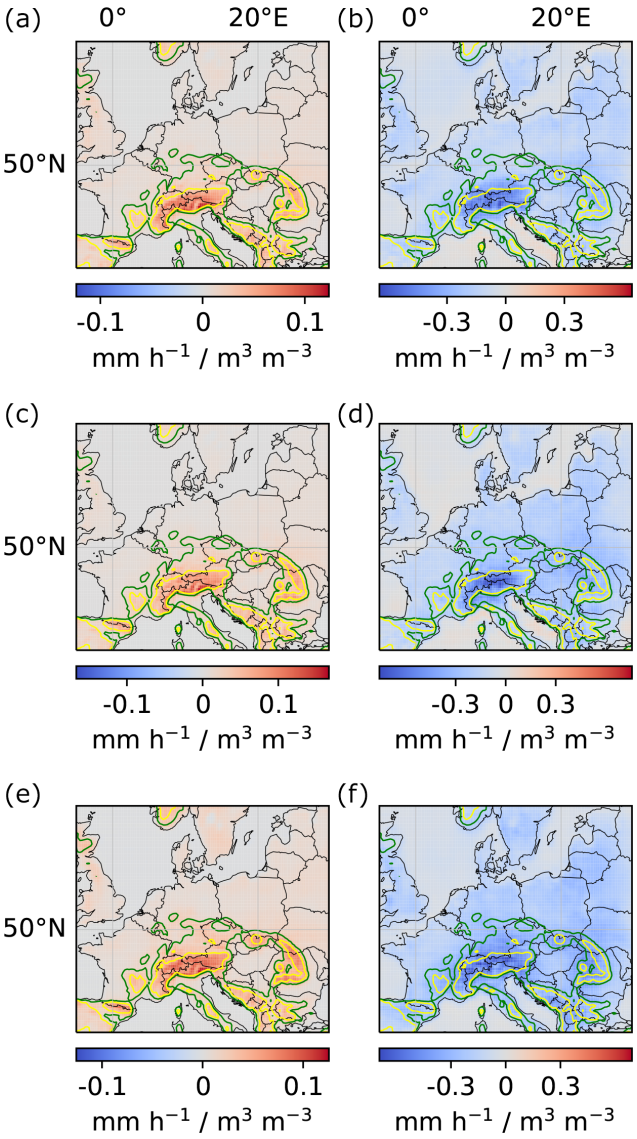

Figure A2. Local and regional soil-moisture–precipitation couplings for models trained only on data from June, July and August, respectively. **(a, c, e)** Local couplings. **(b, d, f)** Regional couplings. **(a, b)** Model trained on data from June. **(c, d)** Model trained on data from July. **(e, f)** Model trained on data from August.

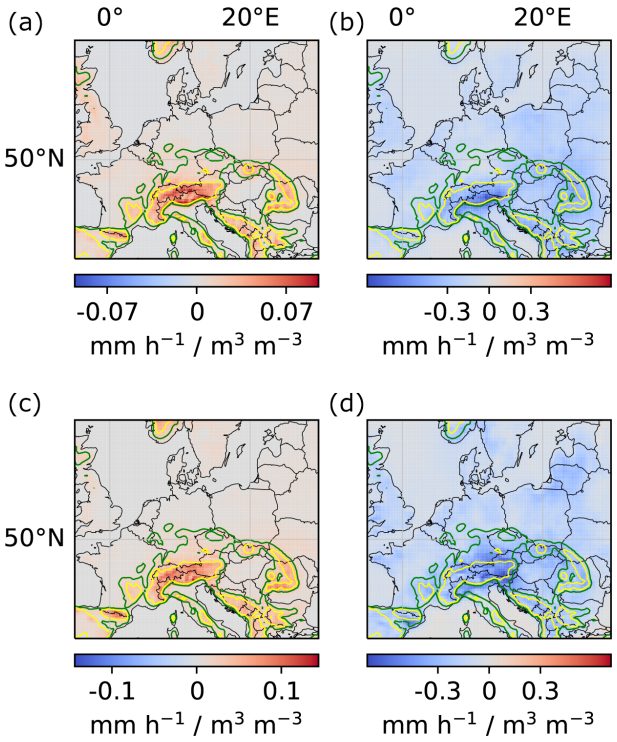

Figure A3. Local and regional soil-moisture–precipitation couplings for models trained on the left and right half of the considered region, respectively. **(a, c)** Local couplings. **(b, d)** Regional couplings. **(a, b)** Model trained on the left half of the considered region. **(c, d)** Model trained on the right half of the considered region (see Appendix Fig. A4). Note that the models were trained only on the left and right half, respectively, but the model architecture allows us to compute local and regional couplings for the entire region.

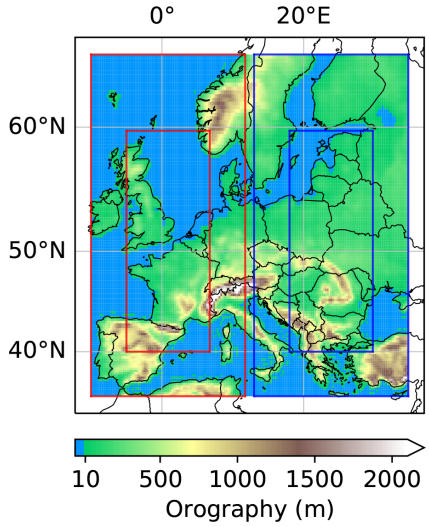

Figure A4. Location variant tasks. The input region was divided into a left and a right input region with corresponding target regions (indicated by the red and blue boxes).

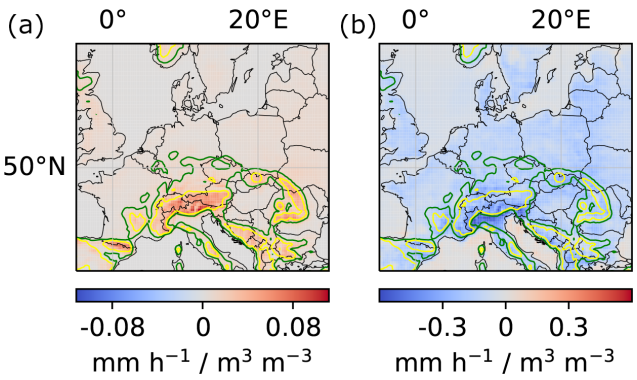

Figure A5. Sum of local and regional soil-moisture–convective-precipitation and soil-moisture–large-scale-precipitation couplings. **(a)** Sum of local couplings. **(b)** Sum of regional couplings. See Appendix Fig. A6 for soil-moisture–convective-precipitation and soil-moisture–large-scale-precipitation couplings.

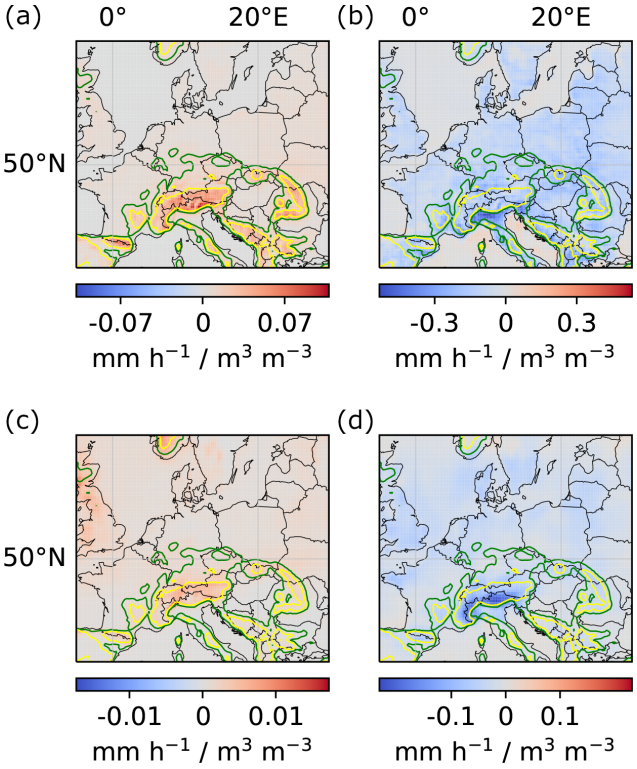

Figure A6. Local and regional soil-moisture–convective-precipitation and soil-moisture–large-scale-precipitation couplings. **(a, c)** Local couplings. **(b, d)** Regional couplings. **(a, b)** Soil-moisture–convective-precipitation coupling. **(c, d)** Soil-moisture–large-scale-precipitation coupling.

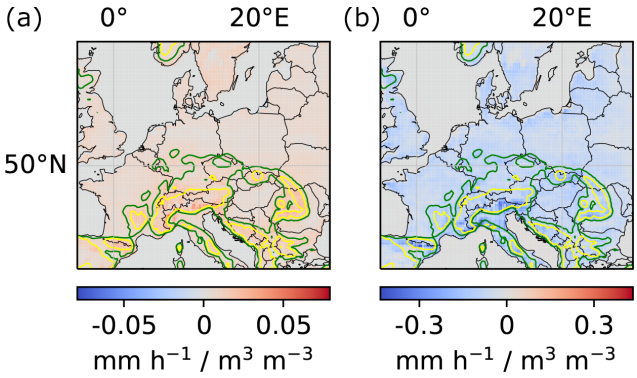

Figure A7. Product of local soil-moisture–evaporation and local and regional evaporation–precipitation couplings. **(a)** Product of local soil-moisture–evaporation and local evaporation–precipitation couplings. **(b)** Product of local soil-moisture–evaporation and regional evaporation–precipitation couplings. See Appendix Fig. A8 for local soil-moisture–evaporation and local and regional evaporation–precipitation couplings.

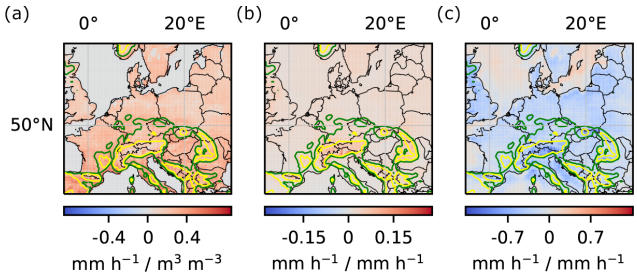

Figure A8. Local soil-moisture–evaporation and local and regional evaporation–precipitation couplings. **(a)** Local soil-moisture–evaporation coupling. **(b)** Local evaporation–precipitation coupling. **(c)** Regional evaporation–precipitation coupling.

Code and data availability. The ERA5 climate reanalysis data (https://doi.org/10.24381/cds.adbb2d47; Hersbach et al., 2023) underlying this study are publicly available. The code to reproduce the study can be found at https://doi.org/10.5281/zenodo.6385040 (Tesch et al., 2022).

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 contact author has declared that none of the authors has any competing interests.

Disclaimer. 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 data (Hersbach et al., 2023) 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.

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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 Initiative and Networking Fund of the Helmholtz Association (HGF) through the project Advanced Earth System Modelling Capacity and the Fraunhofer Cluster of Excellence Cognitive Internet Technologies.

Financial support. This research has been supported by the Helmholtz-RSF Joint Research Group through the project "European hydro-climate extremes: mechanisms, predictability and impacts" (grant no. HRSF-0038).

The article processing charges for this open-access publication were covered by the Forschungszentrum Jülich.

Review statement. This paper was edited by Richard Mills and reviewed by Matthew Knepley and Chaopeng Shen.

References

Adler, B., Kalthoff, N., and Gantner, L.: Initiation of deep convection caused by land-surface inhomogeneities in West Africa: a modelled case study, Meteorol. Atmos. Phys., 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, J. Geophys. Res.-Atmos., 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, Q. J. Roy. Meteor. Soc., 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 (last access: 16 April 2023), 2016.

Ebert-Uphoff, I. and Deng, Y.: Causal discovery in the geosciences – Using synthetic data to learn how to interpret results, Comput. Geosci., 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, B. Am. Meteorol. Soc., 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 Resour. Res., 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, J. Hydrometeorol., 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, J. Hydrometeor., 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, J. Atmos. Sci., 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, Mon. Weather Rev., 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, J. Hydrometeorol., 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), 1–3 October 2018, Turin, Italy, 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, 6443, 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 Ter-

restrial Model, Water Resour. Res., 57, e2020WR027828, 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 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], https://doi.org/10.24381/cds.adbb2d47, 2023.

Hesterberg, T.: What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum, https://arxiv.org/abs/1411.5279 (last access: 16 April 2023), 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, J. Geophys. Res.-Atmos., 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, Geophys. Res. Lett., 44, 11682–11691, 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 (last access: 16 April 2023), 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, J. 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 (last access: 16 April 2023), 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, B. Am. Meteorol. Soc., 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 T. Inform. Theor., 39, 1404–1408, https://doi.org/10.1109/18.243457, 1993.

Molnar, C.: Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/ (last access: 16 April 2023), 2019.

Montavon, G., Samek, W., and Müller, K.: Methods for interpreting and understanding deep neural networks, Digit. Signal Process., 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., Curran Associates, Inc., 8026–8037, http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf (last access: 16 April 2023), 2019.

Pearl, J.: Causal inference in statistics: An overview, Statistics Surveys, 3, 96–146, 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, 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., Springer International Publishing, Cham, 234–241, https://arxiv.org/abs/1505.04597 (last access: 16 April 2023), 2015.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J.: Explainable Machine Learning for Scientific Insights and Discoveries, IEEE Access, 8, 42200–42216, https://doi.org/10.1109/ACCESS.2020.2976199, 2020.

Runge, J.: Causal network reconstruction from time series: From theoretical assumptions to practical estimation, Chaos, 28, 075310, 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, 2553, 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, Proc. 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, B. Am. Meteorol. Soc., 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, Nat. Geosci., 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, J. 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-Sci. Rev., 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, 527–536, AUAI Press, Arlington, Virginia, USA, 2010.

Taylor, C. M.: Detecting soil moisture impacts on convective initiation in Europe, Geophys. Res. Lett., 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, Nat. Geosci., 4, 430–433, https://doi.org/10.1038/ngeo1173, 2011.

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

Tesch, T., Kollet, S., and Garcke, J.: Causal deep learning models for studying the Earth system: soil moisture-precipitation coupling in ERA5 data across Europe – Software Code, Zenodo [code], https://doi.org/10.5281/zenodo.6385040, 2022.

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/ (last access: 16 April 2023), 2017.

Toms, B. A., Barnes, E. A., and Ebert-Uphoff, I.: Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability, J. Adv. Model. Earth Sy., 12, e2019MS002002, 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 Resour. Res., 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?, Geophys. Res. Lett., 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, J. Mach. Learn. Res., 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.

C. Converse local and non-local soil-moisture–precipitation couplings across Europe

Research article ready for submission to a scientific journal.

C.1. Research article

Converse local and non-local soil-moisture–precipitation couplings across Europe

Tobias Tesch[1,2], Stefan Kollet[1,2], Jochen Garcke[3,4], Stergios Kartsios[5], and Eleni Katragkou[5]

[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, 52425 Jülich, Germany
[3]Fraunhofer SCAI, 53757 Sankt Augustin, Germany
[4]Institut für Numerische Simulation, Universität Bonn, 53115 Bonn, Germany
[5]Department of Meteorology and Climatology, School of Geology, Aristotle University of Thessaloniki, Thessaloniki, Greece

Keypoints

- We use causal deep learning models to study soil-moisture–precipitation coupling in reanalysis and simulation data.
- We find a positive impact of local soil moisture changes on local precipitation and a negative impact on non-local precipitation.
- The impact is particularly strong in and around mountainous regions and ridges.

Abstract

Soil moisture affects the temperature and humidity profiles of the atmosphere, thereby influencing the development and onset of precipitation. However, it remains an open question if an increase in soil moisture leads to an increase or decrease in precipitation. Here, we address this question by applying a recently proposed statistical approach of causal deep learning models to ERA5 climate reanalysis data as well as data from a convection-permitting simulation across Europe. In particular, the considered approach accounts for

non-local effects of soil moisture changes on precipitation and the difference between causation and correlation, both commonly being neglected in studies on soil-moisture–precipitation coupling. We find that local increases in soil moisture lead to local increases in precipitation, while decreasing non-local precipitation. In this diverging response, the non-local coupling strength exceeds the local coupling strength. Further, we find soil-moisture–precipitation coupling to be strongest in and around mountainous regions.

Plain Language Summary

It is well-known that soil moisture affects precipitation, but it remains an open question how, i.e. if an increase in soil moisture leads to more or less precipitation. A better understanding of the impact of soil moisture on precipitation may improve weather and climate predictions. Here, we study the impact using a recently proposed statistical approach of causal deep learning models. The approach overcomes several common limitations of previous studies on soil-moisture–precipitation coupling. Considering different data sets, we find that local increases in soil moisture lead to more precipitation locally, but less precipitation in a neighborhood. Moreover, we find that mountains enhance the strength of these effects. A key finding for future studies on soil-moisture–precipitation coupling is the importance of non-local effects, which have mostly been neglected in previous studies.

1 Introduction

In order to improve process understanding, weather and climate predictions and ultimately enhance decision-making capabilities that protect life and property, the study of soil-moisture–precipitation (SM–P) coupling, i.e. the question how soil moisture (SM) affects precipitation, has been ongoing for several decades and remains an active area of research (Liu et al., 2022; Santanello et al., 2018; Seneviratne et al., 2010). The impact of SM on precipitation commences with the impact of SM on the land surface water and energy balances. In these balances, the role of SM is twofold: on the one hand, an increase in SM can increase the amount of available energy, because SM can decrease albedo and surface temperature and thereby reduce outgoing short- and longwave radiation (Eltahir, 1998; Hauck et al., 2011; Schär et al., 1999). On the other hand, an increase in SM can increase the fraction of available energy that is transformed into latent heat flux and decrease the fraction that is transformed into sensible heat flux (Seneviratne et al., 2010).

Multiple pathways for SM–P coupling arise from these effects (see Figure 1). While an increase in latent heat flux may lead to an increase in precipitation via an increase in atmospheric water content (Eltahir, 1998) or via an increase in moist static energy within the boundary layer (Findell and Eltahir, 2003a,b; Gentine et al., 2013), a higher sensible heat flux may lead to stronger thermals and growth of the atmospheric boundary layer, which can induce convective activity and trigger precipitation (Findell and Eltahir, 2003a,b; Gentine et al., 2013; Hohenegger et al., 2009). Lastly, spatial heterogeneity in sensible and latent heat fluxes can cause spatial heterogeneity in the temperature and humidity profiles of the lower atmosphere, which in turn can induce and affect mesoscale circulations and precipitation (Adler et al., 2011; Eltahir, 1998; Gentine et al., 2019; Taylor, 2015; Taylor et al., 2011). These different pathways of SM–P coupling

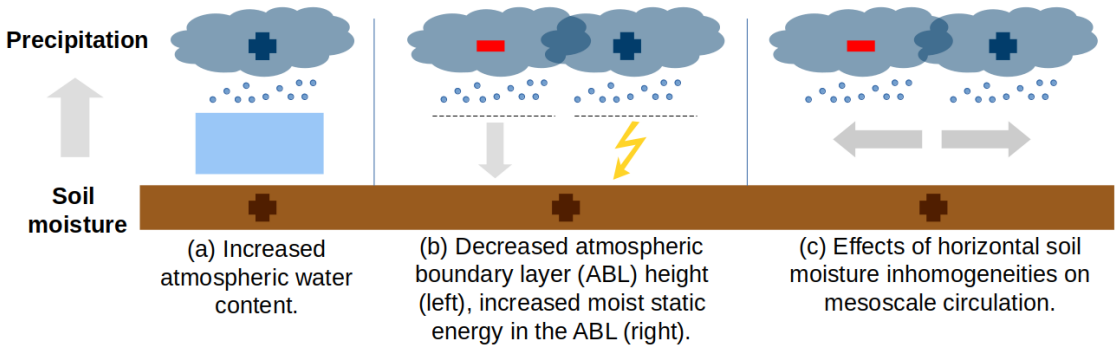

Figure 1: Concurring effects of soil moisture increases on subsequent precipitation. Originally published in (Tesch et al., 2023).

suggest that an increase in SM can lead to both an increase (*positive* coupling) and a decrease (*negative* coupling) in precipitation. Furthermore, they suggest that the impact of an increase in SM on precipitation amount may differ from the impact on precipitation *probability*, i.e. on whether or not a precipitation event of arbitrary magnitude occurs at all (hereafter referred to by soil-moisture–precipitation-*probability* coupling). The different pathways of SM–P coupling make the coupling extremely complex and arguably impossible to determine from theoretical considerations alone.

Over the last decades, many studies have investigated SM–P coupling using different modelling and statistical approaches (Liu et al., 2022; Santanello et al., 2018; Seneviratne et al., 2010). By modelling approaches, we refer to approaches that study SM–P coupling by performing multiple simulations with differing soil moisture conditions, e.g. with different soil moisture initializations. In contrast, by statistical approaches, we refer to analyses of observational data, reanalysis data and analyses of single simulations. From a causality perspective (Runge et al., 2019), modelling approaches correspond to experiments that intervene into the system of interest and evaluate the effects of these interventions. Accordingly, statistical approaches correspond to approaches that learn causal relations from purely observational data.

While most modelling studies based on standard, low-resolution Earth system models have indicated positive SM–P coupling (Seneviratne et al., 2010; Taylor et al., 2012), it has been shown that the coupling is sensitive to the parameterization of convection in low-resolution modelling frameworks to an extent that even the sign of the coupling may be reversed (Hohenegger et al., 2009; Leutwyler et al., 2021; Taylor et al., 2013). While many uncertainties remain (Cioni and Hohenegger, 2017; Kendon et al., 2021), high-resolution (*convection-permitting* (CP)) simulations have been found to agree better with observations (Hohenegger et al., 2009; Leutwyler et al., 2021; Taylor et al., 2013). Several modelling studies based on CP simulations have confirmed the complexity of SM–P coupling (Barthlott and Kalthoff, 2011; Baur et al., 2018; Cioni and Hohenegger, 2017; Hauck et al., 2011; Henneberg et al., 2018; Hohenegger et al., 2009; Imamovic et al., 2017; Leutwyler et al., 2021; Schneider et al., 2019), finding positive and negative coupling depending on the considered region and synoptic situation. Nevertheless, in summary, these modelling studies indicate that SM changes at large scales (in general over the entire simulation domain) on average have a positive impact on the total amount of precipitation, while the impact of SM changes at smaller scales (in parts of the domain) seems to be more involved. Further, they indicate that increases in SM on average have a negative impact on the probability of precipitation events. Note however that computational constraints still

limit the suite of feasible experiments that can be performed at CP resolution, e.g. (Leutwyler et al., 2021) being the only study on SM–P coupling that uses CP simulations at continental scale and spanning several years instead of smaller domains and/or considering case studies on daily time scales.

Statistical approaches for studying SM–P coupling have lower computational costs and can directly be applied to observational data in many cases, circumventing uncertainties due to imperfect representations of SM–P coupling in numerical models. Summarizing the results from several recent studies on SM–P coupling based on statistical approaches (Aires et al., 2014; Findell et al., 2011; Ford et al., 2015, 2018; Froidevaux et al., 2014; Graf et al., 2021; Guillod et al., 2014; Holgate et al., 2019; Li et al., 2020; Tuttle and Salvucci, 2016; Welty and Zeng, 2018), they agree with the findings from modelling studies in that an increase in SM tends to be associated with a (local) increase in precipitation. However, most of these studies also indicate that an increase in SM increases the probability of precipitation events, which is in contrast to the results from the modelling studies. Furthermore, a strong dependence of the coupling sign on the considered data set (Ford et al., 2018; Guillod et al., 2014), the synoptic situation (Ford et al., 2018; Froidevaux et al., 2014; Holgate et al., 2019; Welty and Zeng, 2018), and the considered region (e.g. Aires et al., 2014; Tuttle and Salvucci, 2016) have been reported.

Here, we apply a recently proposed statistical approach for studying complex relations in the Earth system (Tesch et al., 2023) to study SM–P coupling in 41 years of ERA5 climate reanalysis data (Hersbach et al., 2018) across Europe and 15 years of data from a CP simulation across central Europe (Tesch et al., 2022a). Using this approach, we study the impact of SM changes at small spatial scales (*local* SM changes) on both local and non-local precipitation. Note that modelling approaches are not suitable to study the effects of a large number of separate, local SM changes due to computational constraints. On the other hand, previous statistical approaches for studying the effects of SM changes on precipitation have generally neglected non-local effects although they have been shown to be important for overall SM–P coupling (Seneviratne et al., 2010; Wei and Dirmeyer, 2019).

In addition to integrating non-local effects of SM changes, the considered statistical approach integrates insights from causality research (Pearl, 2009) to estimate the causal effect of SM changes on precipitation, rather than mere statistical associations. Previously, this has only been done in (Tuttle and Salvucci, 2016) and the follow-up works (Li et al., 2020; Tuttle and Salvucci, 2017). In particular, our study differs from these studies in that we consider non-local effects of SM on precipitation. Second, to quantify SM–P coupling, we consider the average effect that a small change in SM would have on precipitation, while Tuttle and Salvucci (2016) and follow-up works consider the average difference in precipitation (probability) between “wet” and “dry” days. Third, we consider both SM–precipitation coupling and SM–precipitation-*probability* coupling, while Tuttle and Salvucci (2016) and follow-up works focus on SM–precipitation-probability coupling. Lastly, our approach differs in the input variables that are considered in the statistical model in order to prevent confounding and obtain the actual causal effect of SM changes on precipitation. We formally justify our choice of input variables within the framework of structural causal models (see section 2.2).

2 Data and Method

2.1 Data

We apply the considered statistical approach to ERA5 hourly data (Hersbach et al., 2018), which is a reanalysis of the past decades (1950 to today) provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). ERA5 data contains hourly estimates for a large number of atmospheric, ocean-wave and land-surface quantities on a regular latitude-longitude grid of 0.25 degrees (≈ 30 km). In this study, we consider ERA5 data from 1979 to 2019 across Europe (see region depicted in upper panels of Figure 3).

In addition, we apply the considered statistical approach to data from a convection-permitting (CP) simulation across central Europe (see region depicted in lower panels of Figure 3; Tesch et al., 2022a). The simulation was performed with the WRF model (the Weather Research and Forecasting modeling system, Powers et al., 2017), version 3.8.1, using the Advanced Research dynamical core (Skamarock et al., 2008). It was driven by the ERA-Interim reanalysis and covers the years 2000 to 2014, providing hourly estimates of several Earth system variables on a rotated latitude-longitude grid of 0.0275 degrees (≈ 3 km). The considered simulation is part of the first multi-model ensemble of regional climate simulations at kilometer-scale resolution (Ban et al., 2021; Coppola et al., 2018; Pichelli et al., 2021). More information on the CP simulation can be found in (Ban et al., 2021).

Because SM–P coupling in Europe seems to be strongest during the summer months (e.g. Schär et al., 1999), we only consider data from the months June, July and August. Moreover, we restrict our analyses to daytime processes, considering the effect of SM on precipitation occurring between 11 am and 11 pm UTC. In this study, SM refers to the volumetric soil water in the upper soil layer, which has depths of 7 cm in the ERA5 data and 10 cm in the CP data. Note that previous studies (e.g. Guillod et al., 2015) found SM–P coupling to be insensitive to the considered depth.

2.2 Method

In this work, we study SM–P coupling using a recently proposed statistical approach of causal deep learning models for studying complex relations in the Earth system (Tesch et al., 2023). In order to study SM–P coupling, a *causal* deep learning (DL) model is trained to predict precipitation at each pixel in a target region given SM (and additional input variables; see below) at each pixel in the corresponding input region. A sensitivity analysis of the trained model is performed to analyze how precipitation changes when SM is changed.

A DL model is called *causal* if it approximates the map (Tesch et al., 2023)

$$(\mathrm{SM}[t], \{C_i[t]\}_{i=1}^k) \rightarrow \mathbb{E}\left[\mathrm{P}[t+4\ \mathrm{h}] | do(\mathrm{SM}[t]), \{C_i[t]\}_{i=1}^k\right], \tag{1}$$

where SM$[t]$ represents SM at time t at all input pixels, $\{C_i[t]\}_{i=1}^k$ represent the additional input variables (see below) at the same pixels, and P$[t+4$ h$]$ represents the accumulated precipitation over the time interval $[t+3$ h, $t+4$ h$]$ at all target pixels. The expression $do(\mathrm{SM}[t])$ distinguishes a causal DL model from a standard DL model. It originates from the framework of structural causal models (Pearl, 2009) and represents an

arbitrary intervention into the Earth system. Thus, the term on the right hand side of equation (1) is the expected value of $\mathrm{P}[t + 4\ \mathrm{h}]$ given the variables $\{C_i[t]\}_{i=1}^k$ and given that one intervened into the Earth system and set SM at time t to some arbitrary value $\mathrm{SM}[t]$ (as one could do it in a modelling approach for studying SM–P coupling). The key for obtaining such a causal DL model is a suitable choice of additional input variables $\{C_i[t]\}_{i=1}^k$ to prevent confounding (Tesch et al., 2023). Topography, for example, correlates strongly with soil moisture and precipitation. By including topography as an additional input variable to the DL model, we ensure that the DL model does not erroneously attribute the strong effects of topography on precipitation to soil moisture. For both data sets, we perform two experiments with different choices of input variables summarized in Figure 2. More details on the choice of input variables are given in (Tesch et al., 2023, section 3.3). Further, Supplementary Text 1 contains details on the considered DL models and the training procedure for these models.

In the sensitivity analysis, we do the following: before training the DL models, we partitioned the available data into training, validation and test sets (see Supplementary Text 1). For each time step in the test set (all time steps in the years 1986, 1994, 2002, 2010 and 2018 for the ERA5 data, and all time steps in the years 2005 and 2013 for the CP data) and each tuple (i, j) of two target pixels, we compute the partial derivative of the DL model's precipitation prediction at pixel j with respect to the SM input at pixel i. These derivatives approximate the corresponding derivatives of the map from equation (1), i.e.

$$s_{ij} = \frac{\partial \mathbb{E}\left[\mathrm{P}[t + 4\ \mathrm{h}]_j | do(\mathrm{SM}[t]), \{C_n[t]\}_{n=1}^k\right]}{\partial \mathrm{SM}[t]_i}, \tag{2}$$

which represent how precipitation at pixel j at time $t + 4$ h would change if we intervened into the Earth system and slightly increased SM at pixel i at time t. We average these derivatives over all time steps in the respective test set to obtain the average impact of a slight increase in SM at pixel i on subsequent precipitation at pixel j (hereafter referred to by $\overline{s_{ij}}$). Further, in this study, we focus on two aggregations of these derivatives. Namely, for each target pixel i, we consider $\overline{s_{ii}}$, i.e. the impact of a slight increase in SM at pixel i on subsequent precipitation at pixel i itself, referred to as *local* SM–P coupling, and $\sum_{j \in \{\text{target pixels}\}} \overline{s_{ij}}$, i.e. the impact of a slight increase in SM at pixel i on subsequent precipitation anywhere in the target region, referred to as *regional* SM–P coupling. Note that due to the architecture of the considered DL models, changes in the SM input at some pixel i can only affect the precipitation predictions of the models inside a certain neighborhood (52×52 pixels for the ERA5 data and 116×116 pixels for the CP data), such that $\overline{s_{ij}}$ is zero for all target pixels j outside this neighborhood (see Supplementary Text 1).

To assess the significance of the obtained couplings, we permute the training years such that each original year y_{orig} is mapped randomly to some other year $y_{\text{perm}}(y_{\text{orig}})$ (e.g. 1979 to 1985, 1980 to 2003, ..., 1987 to 1980 and so on). Then, we modify the training data set by replacing the SM input field for each time step t by the SM input field from a corresponding time step t', where t' is the time step corresponding to the same time of the day and day of the year as t, but to the year $y_{\text{perm}}(y_{\text{orig}}(t))$ rather than to year $y_{\text{orig}}(t)$ (illustrated in Supplementary Figure 4). Next, we train a new, randomly initialized instance of the DL model on the modified training set (hereafter referred to as *variant model*). We repeat this procedure 10 times and evaluate whether the mean squared error (MSE) on the original test set is significantly smaller for 10 randomly initialized instances of the DL model trained on the original training set (*original models*) than

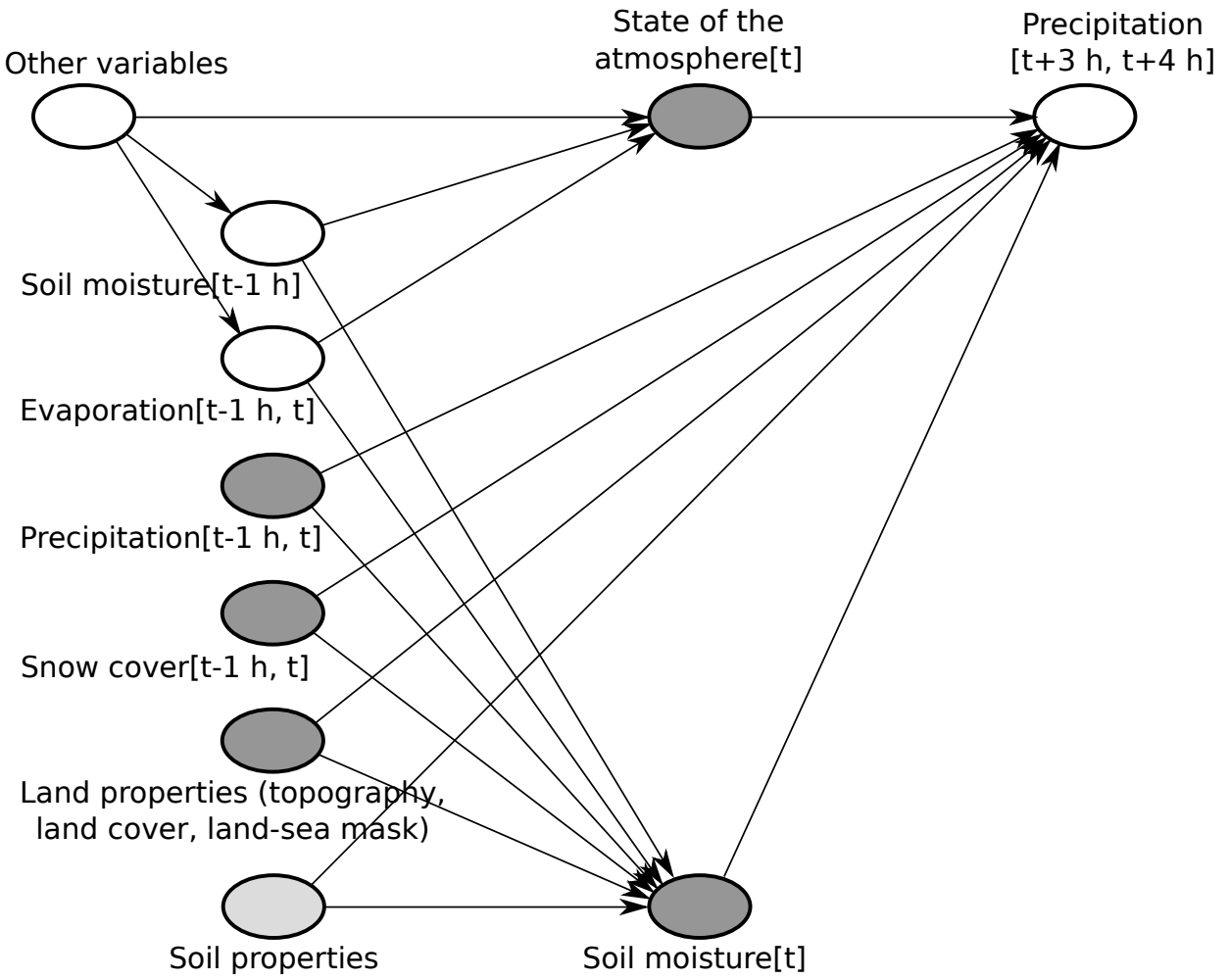

Figure 2: Causal graph summarizing our choice of additional input variables $\{C_i[t]\}_{i=1}^{k}$. Figure adapted from Tesch et al. (2023). Information on the interpretation of the causal graph and the choice of input variables is given in (Tesch et al., 2023, section 2.1 and section 3.3, respectively). The dark grey nodes in the graph represent the chosen input variables. For both data sets, we perform two experiments with varying input variables, where both experiments mainly differ in the approximation of the state of the atmosphere at time t: in the first experiment, we represent this state only by near-surface variables, namely, by near-surface temperature$[t]$ and humidity$[t]$, as well as 100 m U and V components of wind$[t]$. In the second experiment, we additionally include surface pressure$[t]$, vertically integrated atmospheric water content$[t]$ and water vapor content$[t]$ as well as boundary layer height$[t]$ to describe the state of the atmosphere. Further, in all experiments, we include precipitation$[t$ - 1 h$, t]$ and topography as input variables. Snow cover $[t$ - 1 h$, t]$ is included in all experiments except from the first experiment for the ERA5 data for compatibility of this experiment with (Tesch et al., 2023). In both experiments using ERA5 data, we further include high and low vegetation cover as well as land-sea mask, which were not available for the CP data. In addition to these variables, we include short- and long-wave radiation at the land surface$[t]$ in all experiments.

for the 10 variant models. This analysis is related to Granger causality (Granger, 1969) and the analysis in (Tuttle and Salvucci, 2016), where the performance of a model for predicting precipitation given several input variables including SM is compared to the performance of a model with the same input variables but without SM in order to determine whether SM "Granger-causes" precipitation. Using permuted rather than no SM as input for the variant models as described above, our analysis also indicates whether the original models learn information on SM–P coupling apart from noise, and the correlations between SM and topography or seasonality (which are preserved by the described modification of the training set). In addition, we evaluate at what pixels local and regional couplings obtained from the original models differ significantly from the respective couplings obtained from the variant models. For more details on the significance analyses, we refer to Supplementary Text 3.

3 Results and Discussion

In the first experiment for the ERA5 data, the performance of the original models is better than that of the variant models with a confidence of 1. This means that the original models learned useful information on SM–P coupling in terms of predictive performance apart from noise, and correlations between SM and topography or seasonality (see section 2.2). The upper row in Figure 3 shows the local and regional SM–P couplings obtained from these models. Missing hatching indicates where the couplings differ significantly from the couplings obtained from the variant models, i.e. where they not only reflect noise or correlations between SM and topography or seasonality. Most strikingly, we observe opposite signs of local and regional couplings, indicating that an increase in local SM leads to a local increase but an even stronger, non-local decrease in subsequent precipitation. Further, we observe that mountainous regions and ridges enhance the magnitude of soil-moisture–precipitation coupling. Note that this is *not* due to enhanced precipitation in these regions, as ensured by including topography as an input variable, and confirmed by the performed significance analyses (see section 2.2). In the second experiment, where we use more input variables in addition to SM to further reduce the impact of confounding variables on the obtained SM–P couplings (see Figure 2), we find very similar local and regional couplings (see upper row in Supplementary Figure 6). The main difference in the couplings obtained in the second experiment is that the regional coupling is strongest in the vicinity of mountains rather than directly in the mountains (most prominently visible in the Alps).

In the first experiment for the CP data, the performance of the original models is better than that of the variant models with a confidence of 0.9. The obtained local and regional SM–P couplings (see bottom row in Figure 3) are qualitatively similar to the ones obtained for the ERA5 data indicating positive local and negative regional couplings, which are strongest in the vicinity of mountainous regions and ridges. In contrast to the couplings obtained for the ERA5 data, at many pixels the couplings are not significant, i.e. do not differ significantly from the couplings obtained for the variant models. Indeed, we believe that it is more challenging for the DL model to learn SM–P coupling correctly for the high-resolution CP data than for the ERA5 data, because the highly chaotic nature of convection in CP simulations may mask SM–P coupling (Henneberg et al., 2018), and because less training years were available for the CP data than for the ERA5 data, while the considered DL model even had more parameters (see Supplementary Text 1).

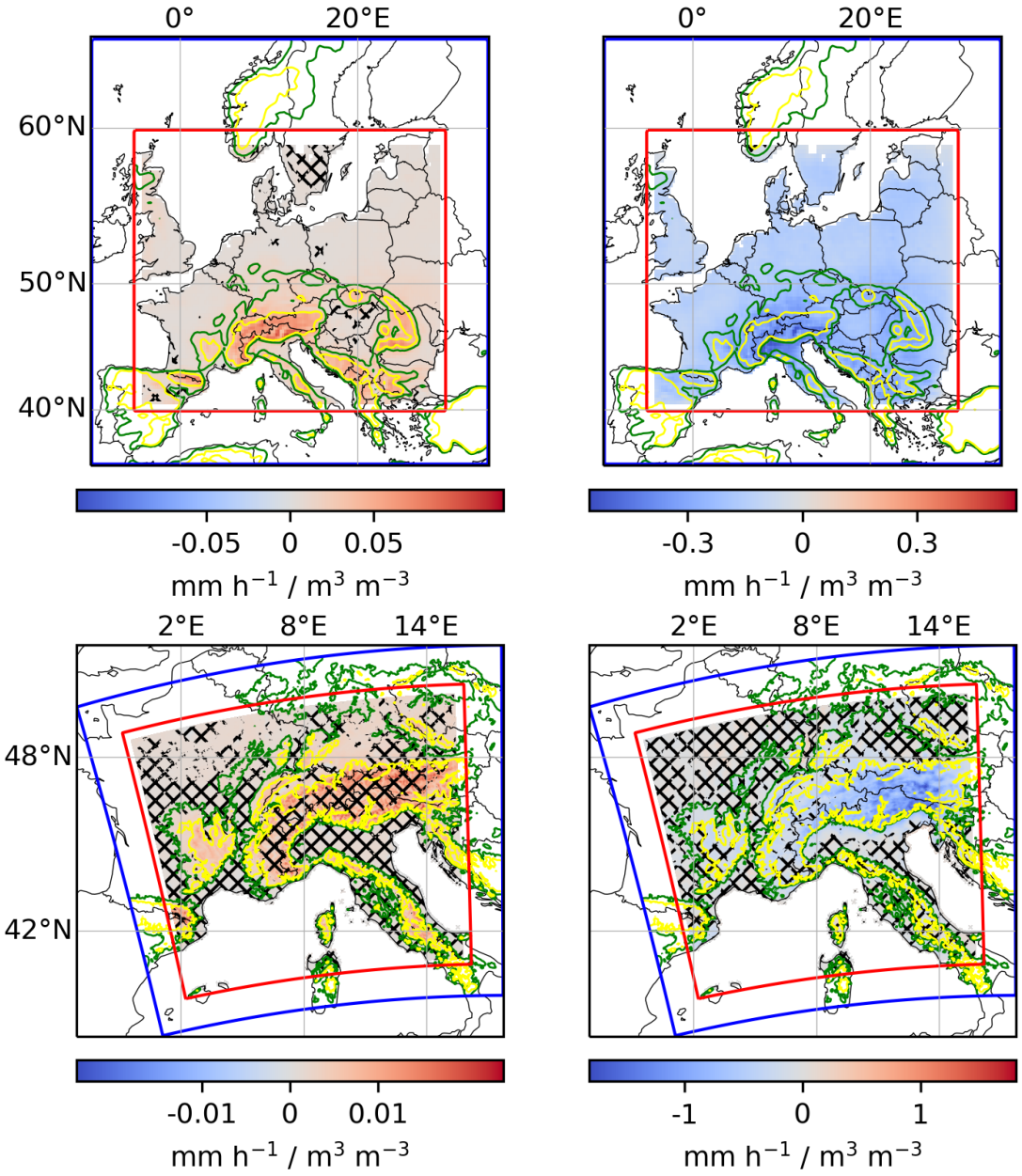

Figure 3: Local (left) and regional (right) SM–P couplings for ERA5 data (top) and CP data (bottom). Local SM–P couplings describe the impact of local SM changes (m^3 water $\cdot$ m^{-3} soil) on local precipitation (mm h^{-1}), while regional SM–P couplings describe the impact of local SM changes on regional precipitation (see section 2.2). For better comparability of the strength of local and regional couplings, the unit mm h^{-1} refers to a single pixel in all panels. Missing hatching at a pixel indicates that the coupling at that pixel does *not* only reflect noise or correlations between SM and topography or seasonality (see section 2.2). The blue and red boxes represent the considered input and target regions, respectively (see section 2.2). The input regions comprise 180×120 pixels for the ERA5 data and 436×484 pixels for the CP data, respectively. The target regions comprise 140×80 pixels for the ERA5 data and 348×396 pixels for the CP data, respectively. The couplings are computed for all target pixels except from very few pixels that are subject to boundary effects (see Supplementary Text 2). The green and yellow elevation contour lines indicate 370 m and 750 m, respectively. Supplementary Figure 1 shows the topography of the regions. This figure shows the results obtained for the first experiment (see Figure 2). Similar results obtained for the second experiment are shown in Supplementary Figure 6.

The overall positive local SM–P coupling found in this study is in line with most previous statistical and modelling studies on SM–P coupling (e.g. Leutwyler et al., 2021; Li et al., 2020, see also Introduction), which mostly found that precipitation increases in regions where SM is increased. To the best of our knowledge, there are no statistical studies on the question how a local increase in SM affects subsequent non-local or regional precipitation (i.e. positively or negatively) and the only modelling study based on CP simulations that addresses this question at a similar spatial scale as our study is (Imamovic et al., 2017). The authors simulated multiple times 5 days with typical European summer day conditions and an initially resting atmosphere over an artificial 256 km $\times$ 256 km domain with a central, Gaussian-shaped mountain with varying heights (0-500 m) and a radius of approximately 30 km. Between their simulations, they varied initial SM heterogeneously at the central mountain from -30 % to +30 % in steps of 10 %, with respect to a reference simulation with domain-wide homogeneous initial SM saturation of 60 % (typical European conditions). In this setting, they found that local precipitation was reduced for a local increase in SM (negative coupling), while non-local precipitation was less affected. This is in contrast to our findings on a positive local and a negative non-local SM–P coupling, although it agrees with an overall (regional) negative SM–P coupling. Note however, that Imamovic et al. (2017) considered a different time scale, and specific atmospheric, topographic, and initial SM conditions. Other modelling studies on SM–P coupling are not suitable to answer the question how a local increase in SM affects regional precipitation at the spatial scales considered here, as variations in initial SM are mostly performed across the entire simulation domain or in rather large subdomains (Henneberg et al., 2018; Leutwyler et al., 2021). In the former cases, mostly positive SM–P coupling was found, while in the latter cases, both positive and negative signs of local and regional SM–P couplings were found.

Concerning the particularly strong local and regional couplings in the vicinity of mountainous regions and ridges, our results agree with findings in (Leutwyler et al., 2021), who simulated 10 summer seasons in continental Europe, each with realistic initial spring SM and with homogeneous perturbations of initial SM saturations by ± 25 %. With an overall positive SM–P coupling, they found the strongest coupling (i.e. the largest differences in precipitation between the runs with high and low initial SM saturations) in the Alpine region. Because differences in evaporation between wet and dry runs were small in the Alpine region, they hypothesized that this is due to more humidity being advected to the Alpine region from neighboring regions in the wet runs. An analysis performed in (Tesch et al., 2023) indicates another hypothesis as to why the magnitude of SM–P coupling is enhanced in mountainous regions. Namely, it indicates that SM–evaporation coupling in these regions is weak, but evaporation-precipitation coupling is particularly strong, the latter potentially being caused by more (topographically induced) vertical air movement (topographic lift) in these regions. Since SM–P coupling is the product of SM–evaporation coupling and evaporation-precipitation coupling, a particularly strong evaporation-precipitation coupling could also explain the particularly strong SM–precipitation coupling in the mountains. Using a statistical approach, Li et al. (2020) found a particularly strong positive local impact of SM on next-day precipitation probability at the leeward slope of the Rocky Mountains. They explained this by water vapor being blocked by the mountains leading to evaporation (and eventually precipitation) being strongly controlled by SM rather than by the horizontal transport of water vapor. In the top right panel of Figure 3, we might observe a similar effect in northern Italy. On the other hand, our findings on particularly strong local and regional couplings in and around mountainous

regions and ridges are in contrast with findings in (Imamovic et al., 2017). Namely, Imamovic et al. (2017) performed the above described simulations with varying heights (0-500 m) of the central mountain and also with homogeneous variations of initial SM, and found a weakening of SM–P coupling for higher mountains, which they explain by mountain-valley circulations and associated convective events dominating over the SM–P feedback for higher mountains.

When considering the impact of changes in SM on precipitation *probability* (by replacing the target variable $\mathrm{P}[t + 4\ \mathrm{h}]$ in section 2.2 by $\mathrm{P}[t + 4\ \mathrm{h}] \geq 1$ mm), we found very similar patterns compared to Figure 3 (see Supplementary Figure 7). In both cases, local coupling is positive and regional coupling negative, and all couplings are strongest in or around mountainous regions and ridges. To some extent, this is surprising, as different processes can be relevant for precipitation occurrence and magnitude. For example, in the experiments described above, Leutwyler et al. (2021) found that a homogeneous increase in SM leads to less convection events due to less thermal circulation, but, at the same time, to more intense events due to larger values of convective available potential energy (CAPE).

4 Conclusion

While being important, previous studies on soil-moisture–precipitation coupling have mostly neglected non-local effects of soil moisture changes on precipitation. Here, we applied a recently proposed, statistical approach of causal deep learning models to study the effects of local soil moisture changes on subsequent local and non-local precipitation in ERA5 climate reanalysis data and in data from a high-resolution, convection-permitting simulation. We found that increases in local soil moisture lead to increases in subsequent local precipitation, but to decreases in non-local precipitation. The impact on non-local precipitation exceeds the local impact, leading to an overall negative regional soil-moisture–precipitation coupling. This stresses the importance of taking into account non-local effects in future studies on soil-moisture–precipitation coupling. We found the magnitude of soil-moisture–precipitation coupling to be enhanced in and around mountainous regions and ridges. Note that this is *not* due to enhanced precipitation in these regions, as ensured by including topography as an input variable, and confirmed by the performed significance analyses (see section 2.2). We also found that the average impact of local soil moisture changes on local and regional precipitation is qualitatively very similar to the average impact on local and regional precipitation *probability*.

5 Open Research

In this work, we use publicly available ERA5 climate reanalysis data (Hersbach et al., 2018) as well as publicly available data from a convection-permitting simulation (Tesch et al., 2022a). Software code to reproduce this study is available on Zenodo under the MIT license (Tesch et al., 2022b).

6 Acknowledgements

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'. EK and SKa acknowledge the GRNET HPC-ARIS infrastructure (projects pr003005 and pr009020_thin) and the AUTH-IT scientific center for their support. 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 data (Hersbach 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

B. Adler, N. Kalthoff, and L. Gantner. Initiation of deep convection caused by land-surface inhomogeneities in West Africa: a modelled case study. *Meteorology and Atmospheric Physics*, 112(1-2):15–27, 2011. doi: 10.1007/s00703-011-0131-2.

F. Aires, P. Gentine, K. L. Findell, B. R. Lintner, and C. Kerr. Neural Network–Based Sensitivity Analysis of Summertime Convection over the Continental United States. *Journal of Climate*, 27(5):1958–1979, 2014. doi: 10.1175/jcli-d-13-00161.1.

N. Ban, C. Caillaud, E. Coppola, E. Pichelli, S. Sobolowski, M. Adinolfi, B. Ahrens, A. Alias, I. Anders, S. Bastin, D. Belušić, S. Berthou, E. Brisson, R. M. Cardoso, S. C. Chan, O. B. Christensen, J. Fernández, L. Fita, T. Frisius, G. Gašparac, F. Giorgi, K. Goergen, J. E. Haugen, Ø. Hodnebrog, S. Kartsios, E. Katragkou, E. J. Kendon, K. Keuler, A. Lavin-Gullon, G. Lenderink, D. Leutwyler, T. Lorenz, D. Maraun, P. Mercogliano, J. Milovac, H.-J. Panitz, M. Raffa, A. R. Remedio, C. Schär, P. M. M. Soares, L. Srnec, B. M. Steensen, P. Stocchi, M. H. Tölle, H. Truhetz, J. Vergara-Temprado, H. de Vries, K. Warrach-Sagi, V. Wulfmeyer, and M. J. Zander. The first multi-model ensemble of regional climate simulations at kilometer-scale resolution, part I: evaluation of precipitation. *Climate Dynamics*, 57(1-2):275–302, 2021. doi: 10.1007/s00382-021-05708-w. The considered simulation has the group abbreviation "AUTH".

C. Barthlott and N. Kalthoff. A Numerical Sensitivity Study on the Impact of Soil Moisture on Convection-Related Parameters and Convective Precipitation over Complex Terrain. *Journal of the Atmospheric Sciences*, 68(12):2971–2987, 2011. doi: 10.1175/jas-d-11-027.1.

F. Baur, C. Keil, and G. C. Craig. 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(717):2863–2875, 2018. doi: 10.1002/qj.3415.

G. Cioni and C. Hohenegger. Effect of Soil Moisture on Diurnal Convection and Precipitation in Large-Eddy Simulations. *Journal of Hydrometeorology*, 18(7):1885–1903, 2017. doi: 10.1175/jhm-d-16-0241.1.

E. Coppola, S. Sobolowski, E. Pichelli, F. Raffaele, B. Ahrens, I. Anders, N. Ban, S. Bastin, M. Belda, D. Belusic, A. Caldas-Alvarez, R. M. Cardoso, S. Davolio, A. Dobler, J. Fernandez, L. Fita, Q. Fumiere, F. Giorgi, K. Goergen, I. Güttler, T. Halenka, D. Heinzeller, Ø. Hodnebrog, D. Jacob, S. Kartsios, E. Katragkou, E. Kendon, S. Khodayar, H. Kunstmann, S. Knist, A. Lavín-Gullón, P. Lind, T. Lorenz, D. Maraun, L. Marelle, E. van Meijgaard, J. Milovac, G. Myhre, H.-J. Panitz, M. Piazza, M. Raffa, T. Raub, B. Rockel, C. Schär, K. Sieck, P. M. M. Soares, S. Somot, L. Srnec, P. Stocchi, M. H. Tölle, H. Truhetz, R. Vautard, H. de Vries, and K. Warrach-Sagi. A first-of-its-kind multi-model convection permitting ensemble for investigating convective phenomena over Europe and the Mediterranean. *Climate Dynamics*, 55(1-2): 3–34, 2018. doi: 10.1007/s00382-018-4521-8.

E. A. B. Eltahir. A Soil Moisture–Rainfall Feedback Mechanism: 1. Theory and observations. *Water Resources Research*, 34(4):765–776, 1998. doi: https://doi.org/10.1029/97WR03499.

K. L. Findell and E. A. B. Eltahir. Atmospheric Controls on Soil Moisture–Boundary Layer Interactions. Part I: Framework Development. *Journal of Hydrometeorology*, 4(3):552–569, 2003a. doi: 10.1175/1525-7541(2003)004<0552:acosml>2.0.co;2.

K. L. Findell and E. A. B. Eltahir. Atmospheric Controls on Soil Moisture–Boundary Layer Interactions. Part II: Feedbacks within the Continental United States. *Journal of Hydrometeorology*, 4(3):570–583, 2003b. doi: 10.1175/1525-7541(2003)004<0570:acosml>2.0.co;2.

K. L. Findell, P. Gentine, B. R. Lintner, and C. Kerr. Probability of afternoon precipitation in eastern United States and Mexico enhanced by high evaporation. *Nature Geoscience*, 4(7):434–439, 2011. doi: 10.1038/ngeo1174.

T. W. Ford, A. D. Rapp, S. M. Quiring, and J. Blake. Soil moisture–precipitation coupling: observations from the Oklahoma Mesonet and underlying physical mechanisms. *Hydrology and Earth System Sciences*, 19(8):3617–3631, 2015. doi: 10.5194/hess-19-3617-2015.

T. W. Ford, S. M. Quiring, B. Thakur, R. Jogineedi, A. Houston, S. Yuan, A. Kalra, and N. Lock. Evaluating Soil Moisture–Precipitation Interactions Using Remote Sensing: A Sensitivity Analysis. *Journal of Hydrometeorology*, 19(8):1237–1253, 2018. doi: 10.1175/jhm-d-17-0243.1.

P. Froidevaux, L. Schlemmer, J. Schmidli, W. Langhans, and C. Schär. Influence of the Background Wind on the Local Soil Moisture–Precipitation Feedback. *Journal of the Atmospheric Sciences*, 71(2):782–799, 2014. doi: 10.1175/jas-d-13-0180.1.

P. Gentine, A. A. M. Holtslag, F. D'Andrea, and M. Ek. Surface and Atmospheric Controls on the Onset of Moist Convection over Land. *Journal of Hydrometeorology*, 14(5):1443–1462, 2013. doi: 10.1175/jhm-d-12-0137.1.

P. Gentine, A. Massmann, B. R. Lintner, S. H. Alemohammad, R. Fu, J. K. Green, D. Kennedy, and J. Vilà-Guerau de Arellano. Land–atmosphere interactions in the tropics – a review. *Hydrology and Earth System Sciences*, 23(10):4171–4197, 2019. doi: 10.5194/hess-23-4171-2019.

M. Graf, J. Arnault, B. Fersch, and H. Kunstmann. Is the soil moisture precipitation feedback enhanced by heterogeneity and dry soils? A comparative study. *Hydrological Processes*, 35(9), 2021. doi: 10.1002/hyp.14332.

C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, 1969. doi: 10.2307/1912791.

B. P. Guillod, B. Orlowsky, D. Miralles, A. J. Teuling, P. D. Blanken, N. Buchmann, P. Ciais, M. Ek, K. L. Findell, P. Gentine, B. R. Lintner, R. L. Scott, B. V. den Hurk, and S. I. Seneviratne. Land-surface controls on afternoon precipitation diagnosed from observational data: uncertainties and confounding factors. *Atmospheric Chemistry and Physics*, 14(16):8343–8367, 2014. doi: 10.5194/acp-14-8343-2014.

B. P. Guillod, B. Orlowsky, D. G. Miralles, A. J. Teuling, and S. I. Seneviratne. Reconciling spatial and temporal soil moisture effects on afternoon rainfall. *Nature Communications*, 6(1), 2015. doi: 10.1038/ncomms7443.

C. Hauck, C. Barthlott, L. Krauss, and N. Kalthoff. Soil moisture variability and its influence on convective precipitation over complex terrain. *Quarterly Journal of the Royal Meteorological Society*, 137(S1):42–56, 2011. doi: 10.1002/qj.766.

O. Henneberg, F. Ament, and V. Grützun. Assessing the uncertainty of soil moisture impacts on convective precipitation using a new ensemble approach. *Atmospheric Chemistry and Physics*, 18(9):6413–6425, 2018. doi: 10.5194/acp-18-6413-2018.

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

T. Hesterberg. What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. 2014. doi: 10.48550/arXiv.1411.5279.

C. Hohenegger, P. Brockhaus, C. S. Bretherton, and C. Schär. The Soil Moisture–Precipitation Feedback in Simulations with Explicit and Parameterized Convection. *Journal of Climate*, 22(19):5003–5020, 2009. doi: 10.1175/2009jcli2604.1.

C. M. Holgate, A. I. J. M. V. Dijk, J. P. Evans, and A. J. Pitman. The Importance of the One-Dimensional Assumption in Soil Moisture - Rainfall Depth Correlation at Varying Spatial Scales. *Journal of Geophysical Research: Atmospheres*, 124(6):2964–2975, 2019. doi: 10.1029/2018jd029762.

A. Imamovic, L. Schlemmer, and C. Schär. Collective Impacts of Orography and Soil Moisture on the Soil Moisture-Precipitation Feedback. *Geophysical Research Letters*, 44(22):11,682–11,691, 2017. doi: https://doi.org/10.1002/2017GL075657.

E. J. Kendon, A. F. Prein, C. A. Senior, and A. Stirling. Challenges and outlook for convection-permitting climate modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2195):20190547, 2021. doi: 10.1098/rsta.2019.0547.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. 2017. doi: 10.48550/arXiv.1412.6980.

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

L. Li, W. Shangguan, Y. Deng, J. Mao, J. Pan, N. Wei, H. Yuan, S. Zhang, Y. Zhang, and Y. Dai. A Causal Inference Model Based on Random Forests to Identify the Effect of Soil Moisture on Precipitation. *Journal of Hydrometeorology*, 21(5):1115 – 1131, 2020. doi: 10.1175/JHM-D-19-0209.1.

W. Liu, Q. Zhang, C. Li, L. Xu, and W. Xiao. The influence of soil moisture on convective activity: a review. *Theoretical and Applied Climatology*, April 2022. doi: 10.1007/s00704-022-04046-z.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc., 2019. doi: 10.48550/arXiv.1912.01703.

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

E. Pichelli, E. Coppola, S. Sobolowski, N. Ban, F. Giorgi, P. Stocchi, A. Alias, D. Belušić, S. Berthou, C. Caillaud, R. M. Cardoso, S. Chan, O. B. Christensen, A. Dobler, H. de Vries, K. Goergen, E. J. Kendon, K. Keuler, G. Lenderink, T. Lorenz, A. N. Mishra, H.-J. Panitz, C. Schär, P. M. M. Soares, H. Truhetz, and J. Vergara-Temprado. The first multi-model ensemble of regional climate simulations at kilometer-scale resolution part 2: historical and future simulations of precipitation. *Climate Dynamics*, 56(11-12): 3581–3602, 2021. doi: 10.1007/s00382-021-05657-4.

J. G. Powers, J. B. Klemp, W. C. Skamarock, C. A. Davis, J. Dudhia, D. O. Gill, J. L. Coen, D. J. Gochis, R. Ahmadov, S. E. Peckham, G. A. Grell, J. Michalakes, S. Trahan, S. G. Benjamin, C. R. Alexander, G. J. Dimego, W. Wang, C. S. Schwartz, G. S. Romine, Z. Liu, C. Snyder, F. Chen, M. J. Barlage, W. Yu, and M. G. Duda. The Weather Research and Forecasting Model: Overview, System Efforts, and Future Directions. *Bulletin of the American Meteorological Society*, 98(8):1717–1737, 2017. doi: 10.1175/bams-d-15-00308.1.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and*

Computer-Assisted Intervention – MICCAI 2015, pages 234–241, Cham, 2015. Springer International Publishing. doi: 10.48550/arXiv.1505.04597.

J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, E. H. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler. Inferring causation from time series in Earth system sciences. *Nature Communications*, 10(1), 2019. doi: 10.1038/s41467-019-10105-3.

J. A. Santanello, P. A. Dirmeyer, C. R. Ferguson, K. L. Findell, A. B. Tawfik, A. Berg, M. Ek, P. Gentine, B. P. Guillod, C. van Heerwaarden, J. Roundy, and V. Wulfmeyer. Land–Atmosphere Interactions: The LoCo Perspective. *Bulletin of the American Meteorological Society*, 99(6):1253–1272, 2018. doi: 10.1175/bams-d-17-0001.1.

C. Schär, D. Lüthi, U. Beyerle, and E. Heise. The Soil–Precipitation Feedback: A Process Study with a Regional Climate Model. *Journal of Climate*, 12(3):722–741, 1999. doi: 10.1175/1520-0442(1999)012<0722:tspfap>2.0.co;2.

L. Schneider, C. Barthlott, C. Hoose, and A. I. Barrett. Relative impact of aerosol, soil moisture, and orography perturbations on deep convection. *Atmospheric Chemistry and Physics*, 19(19):12343–12359, 2019. doi: 10.5194/acp-19-12343-2019.

S. I. Seneviratne, T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling. Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, 99(3-4):125–161, 2010. doi: 10.1016/j.earscirev.2010.02.004.

W. Skamarock, J. Klemp, J. Dudhia, D. Gill, D. Barker, W. Wang, X.-Y. Huang, and M. Duda. A Description of the Advanced Research WRF Version 3. 2008. doi: 10.5065/D68S4MVH. Technical report.

C. M. Taylor. Detecting soil moisture impacts on convective initiation in Europe. *Geophysical Research Letters*, 42(11):4631–4638, 2015. doi: 10.1002/2015gl064030.

C. M. Taylor, A. Gounou, F. Guichard, P. P. Harris, R. J. Ellis, F. Couvreux, and M. D. Kauwe. Frequency of Sahelian storm initiation enhanced over mesoscale soil-moisture patterns. *Nature Geoscience*, 4(7): 430–433, 2011. doi: https://doi.org/10.1038/ngeo1173.

C. M. Taylor, R. A. M. de Jeu, F. Guichard, P. P. Harris, and W. A. Dorigo. Afternoon rain more likely over drier soils. *Nature*, 489(7416):423–426, 2012. doi: 10.1038/nature11377.

C. M. Taylor, C. E. Birch, D. J. Parker, N. Dixon, F. Guichard, G. Nikulin, and G. M. S. Lister. Modeling soil moisture-precipitation feedback in the Sahel: Importance of spatial scale versus convective parameterization. *Geophysical Research Letters*, 40(23):6213–6218, 2013. doi: 10.1002/2013gl058511.

T. Tesch, S. Kollet, J. Garcke, E. Katragkou, and S. Kartsios. Data set of the manuscript 'Opposite signs in local and nonlocal soil moisture-precipitation couplings across Europe'. 2022a. doi: 10.26165/juelich-data/YO3JCM.

T. Tesch, S. Kollet, J. Garcke, E. Katragkou, and S. Kartsios. Opposite signs in local and nonlocal soil moisture-precipitation couplings across Europe–Software code. 2022b. doi: 10.5281/zenodo.7034237.

T. Tesch, S. Kollet, and J. Garcke. Causal deep learning models for studying the Earth system. *Geoscientific Model Development*, 16(8):2149–2166, 2023. doi: 10.5194/gmd-16-2149-2023.

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

S. Tuttle and G. Salvucci. Empirical evidence of contrasting soil moisture-precipitation feedbacks across the United States. *Science*, 352(6287):825–828, 2016. doi: 10.1126/science.aaa7185.

S. E. Tuttle and G. D. Salvucci. Confounding factors in determining causal soil moisture-precipitation feedback. *Water Resources Research*, 53(7):5531–5544, 2017. doi: https://doi.org/10.1002/2016WR019869.

J. Wei and P. A. Dirmeyer. Sensitivity of land precipitation to surface evapotranspiration: a nonlocal perspective based on water vapor transport. *Geophysical Research Letters*, 46(21):12588–12597, 2019. doi: 10.1029/2019gl085613.

J. Welty and X. Zeng. Does Soil Moisture Affect Warm Season Precipitation Over the Southern Great Plains? *Geophysical Research Letters*, 45(15):7866–7873, 2018. doi: 10.1029/2018gl078598.

C.2. Supporting information

Supporting Information for "Converse local and non-local soil-moisture–precipitation couplings across Europe"

Tobias Tesch[1,2], Stefan Kollet[1,2], Jochen Garcke[3,4], Stergios Kartsios[5], and Eleni Katragkou[5]

[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, 52425 Jülich, Germany
[3]Fraunhofer SCAI, 53757 Sankt Augustin, Germany
[4]Institut für Numerische Simulation, Universität Bonn, 53115 Bonn, Germany
[5]Department of Meteorology and Climatology, School of Geology, Aristotle University of Thessaloniki, Thessaloniki, Greece

Contents of this file

1. Text S1 to S3
2. Figures S1 to S8
3. Tables S1 to S2

Introduction

In the following sections we present more details on the considered methodology, namely on

1. the considered DL models and training procedure,
2. computational details of the sensitivity analysis, and
3. the significance analyses.

Text S1 – DL Models and Training Procedure

Figure S2 shows the architecture of the considered DL models. As in (Tesch et al., 2023), we chose convolutional neural networks (CNNs) whose architecture was inspired by the U-Net architecture (Ronneberger et al., 2015). An important concept in the context of this architecture is that of receptive fields. Namely, the prediction of the model at some target pixel is fully determined by the input variables in a certain neighborhood, the so called receptive field. For the ERA5 data, the size of the receptive field is $\leq 52 \times 52$ pixels, i.e. the precipitation prediction at a target pixel is fully determined by the input variables in a $\leq 52 \times 52$ pixels neighborhood (see Text S2). For the CP data, a receptive field of $\leq 52 \times 52$ might not be large enough for the model to take into account all relations between SM and precipitation at the considered time scale due to the higher spatial resolution. Therefore, for the CP data, we increased the depth of the DL model (see Figure S2), resulting in a receptive field of $\leq 116 \times 116$ pixels (see Text S2).

Concerning the training of the DL models, we followed the same procedure as described in (Tesch et al., 2023). First, we split the data into training, validation and test sets. For the ERA5 data, the test set comprises the years 1986, 1994, 2002, 2010 and 2018, the validation set comprises the years 1982, 1990, 1998, 2006 and 2014, and the training set comprises the remaining 31 years between 1979 and 2019. For the CP data, the test set comprises the years 2005 and 2013, the validation set comprises the years 2001 and 2009 and the training set comprises the remaining 11 years between 2000 and 2014. After partitioning the data into training, validation and test sets, we used the Adam optimizer (Kingma and Ba, 2017) to adapt the randomly initialized weights of the respective DL model to minimize the mean squared error (MSE) on the respective training set. The validation set was used for early stopping (i.e. we stopped training if the MSE evaluated on the validation set did not improve by some threshold for a certain number of epochs), while the test set was held out during the entire training and previous tuning process of the model.

In terms of implementation, we used 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 to 20 epochs. During training, we further used data augmentation as in (Tesch et al., 2023). Namely, we randomly rotated by 180° (or not) and subsequently horizontally flipped (or not) the considered region for each training sample and each training epoch independently. Further, for the CP data, we randomly cropped the input region from 440×490 pixels to 436×484 pixels.

Text S2 – Sensitivity Analysis

Computing the partial derivatives in equation (2) from the main article is computationally expensive because they have to be computed separately for each target pixel. This is because the DL framework used in this work (Pytorch; Paszke et al., 2019) only implements backpropagation for scalars and not for vectors. When considering the CP data, there are 348×396 target pixels, such that computing the partial derivatives of the model's predictions with respect to the SM input at the 436×484 input pixels for a batch $x \in \mathbb{R}^{N \times 10 \times 436 \times 484}$ of N input time steps (with 10 being the number of input variables) requires a single forward pass of the batch through the DL model and $348 \cdot 396 = 137,808$ backward passes. For the considered DL model, this

is computationally infeasible.

In the following, we detail the strategy that we used to reduce the compute time. First, we noted that the prediction of the DL model at a target pixel is fully determined by the input variables in a certain neighborhood (the *receptive field* of that target pixel). For the DL model considered for the CP data (see Figure S2), the receptive field is smaller or equal to 116×116 pixels (see below). The partial derivatives of the prediction at a target pixel with respect to SM at input pixels outside of the receptive field is 0. When only the input variables in the receptive field are fed to the DL model, the model produces an output of size smaller or equal to 28×28 pixels. One of these pixels corresponds to the originally considered target pixel, while the other pixels are artifacts due to boundary effects (see below). To reduce the compute time, we only pass the input variables in the receptive field of a target pixel to the DL model and backpropagate the prediction at the pixel corresponding to the originally considered target pixel through the model. In this way, computing the partial derivatives of the model's predictions with respect to the SM input at the 436×484 input pixels for a batch $x \in \mathbb{R}^{N \times 10 \times 436 \times 484}$ requires 348×396 forward passes and 348×396 backward passes, but each for an input tensor of shape smaller or equal to $N \times 10 \times 116 \times 116$. In a test with $N = 1$, this was approximately 15 times faster than the naive way of computing the derivatives described above. When the objective is to compute the partial derivatives for many input time steps (in our case the number of input time steps for the sensitivity analysis, i.e. the numer of time steps in the test set of the CP data, is 2208), we expect the speed-up to be even larger, as the reduction in input pixels allows to increase the batch size without causing memory issues.

In the following, we provide the required details for the strategy described above. The architecture of the considered DL model is illustrated in Figure S2. The model contains three types of layers that affect the receptive field of a target pixel, namely convolutional layers (except from the last, 1×1 convolutional layer, which does not affect the receptive field), max-pooling layers, and transposed-convolutional layers. Figure S3 illustrates the effect of these layers on the receptive field in the 1-dimensional analogue of the configurations used in this work (note that one could also use different configurations, for example convolutional kernels of different size, different padding and so on, resulting in different effects on the receptive fields). For a convolutional layer, the output at index $i \in \{0, \dots, n-3\}$ is affected by the input at indexes $i, i+1$ and $i+2$. For a max-pooling layer, the output at index $i \in \{0, \dots, \frac{n}{2} - 1\}$ is affected by the input at indexes $2i$ and $2i+1$. Finally, for a transposed-convolutional layer, the output at index $i \in \{0, \dots, 2n-1\}$ is affected by the input at indexes $\frac{i}{2} - 1$ and $\frac{i}{2}$ if i is even and $\lfloor \frac{i}{2} \rfloor$ and $\lfloor \frac{i}{2} \rfloor + 1$ otherwise (except from the first and the last output indexes, which are affected by boundary effects).

Iterating these dependencies backwards through the architecture of the DL model, we found the receptive field of each target pixel with respect to the input pixels. As there are three transposed-convolutional layers in the model, we had to distinguish between eight different cases (target index $i \bmod 8$). The results are summarized in the first three columns of Table S1. From these considerations, it also follows that the first seven and the last eleven target indexes are subject to boundary effects. Consequently, they were neglected in this work.

When only feeding the input variables in the receptive field of a target index to the considered DL model, the model produces 20 output indexes in cases with a receptive field of size 108, and 28 output indexes in cases with a receptive field of size 116. The output at one of these indexes corresponds to the original

output at the considered target pixel. The other output indexes are due to boundary effects occurring in the transposed-convolutional layers. The easiest way to determine which of the output indexes in this *small output* corresponds to the respective target index in the *original output* is brute force. Namely, it is only necessary to evaluate at what output index in the small output the prediction of the model is identical to the original output at the considered target pixel. To that purpose, we passed several random input fields to the model. The results of this analysis are summarized in column 4 of Table S1.

Note that in the case of the ERA5 data, where there are less target pixels and the DL model consists of less layers, the described procedure for performing the sensitivity analysis also leads to a large speed-up compared to the naive procedure. Table S2 contains the same information as Table S1, but for the smaller model considered for the ERA5 data. In this case, the first and the last three target indexes are subject to boundary effects and neglected.

Text S3 – Significance Analyses

To determine the confidence with that the MSE on the original test set is smaller for the 10 original models than for the 10 variant models, we used a permutation test (Hesterberg, 2014). The procedure was as follows: first, we computed the difference between the mean MSE of the original models and the variant models, $\text{diff}_{\text{orig}}$. Then, we randomly permuted all 20 MSE values and computed the difference between the mean of the first 10 and the mean of the last 10 values. We repeated this $100,000$ times and determined the number of permutations for that the computed difference was larger than $\text{diff}_{\text{orig}}$. This number divided by the total number of considered permutations is the confidence with that the MSE on the original test set is smaller for the original models than for the variant models.

Note that we also compared the pixel-wise MSE of the original models to that of the variant models in order to evaluate whether we could partition the target region into pixels where the MSE of the original models is lower than that of the variant models (e.g. pixels in regions with strong SM–P coupling), and pixels where the MSE of the original models is similar to that of the variant models (e.g. pixels in regions with no or very weak SM–P coupling). However, we obtained chaotic patterns showing increases in MSE at some pixels, but also decreases at other pixels (see Figure S5). We evaluated how this pattern evolved when only one, two, three, four or all five test years were considered in the computation of the MSE and observed that the fraction of pixels showing lower MSE for the original models than for the variant models increased with an increasing number of considered test years. Therefore, we concluded that there are not enough test years to perform this analysis on a pixel-wise level: for the considered number of test years, on a pixel-wise level, the chaotic nature of convection seems to mask the small improvements in precipitation prediction that are due to the inclusion of correct SM data.

To evaluate at what pixels local and regional couplings obtained from the original models differ significantly from the respective couplings obtained from the variant models, we did the following: for each target pixel i, we computed the mean over the local and regional couplings obtained from the ten original models, referred to as $m^{\text{orig}}(l_i)$ and $m^{\text{orig}}(r_i)$, respectively. Next, we computed the mean and the standard deviation over the local and regional couplings obtained from the ten variant models, referred to as $m^{\text{var}}(l_i)$ and $s^{\text{var}}(l_i)$, and

$m^{\text{var}}(r_i)$ and $s^{\text{var}}(r_i)$, respectively. Finally, we evaluated if

$$\left|m^{\text{orig}}(l_i) - m^{\text{var}}(l_i)\right| > 1.645 \cdot s^{\text{var}}(l_i),$$

and

$$\left|m^{\text{orig}}(r_i) - m^{\text{var}}(r_i)\right| > 1.645 \cdot s^{\text{var}}(r_i),$$

respectively. If this was the case, we concluded that the local and regional couplings, respectively, at pixel i obtained from the original models differ significantly from the respective couplings obtained from the variant models. If the values of local and regional couplings obtained from the variant models were normally distributed, this procedure would test whether $m^{\text{orig}}(l_i)$ and $m^{\text{orig}}(r_i)$, respectively, lie within the central 90 % of the respective distributions.

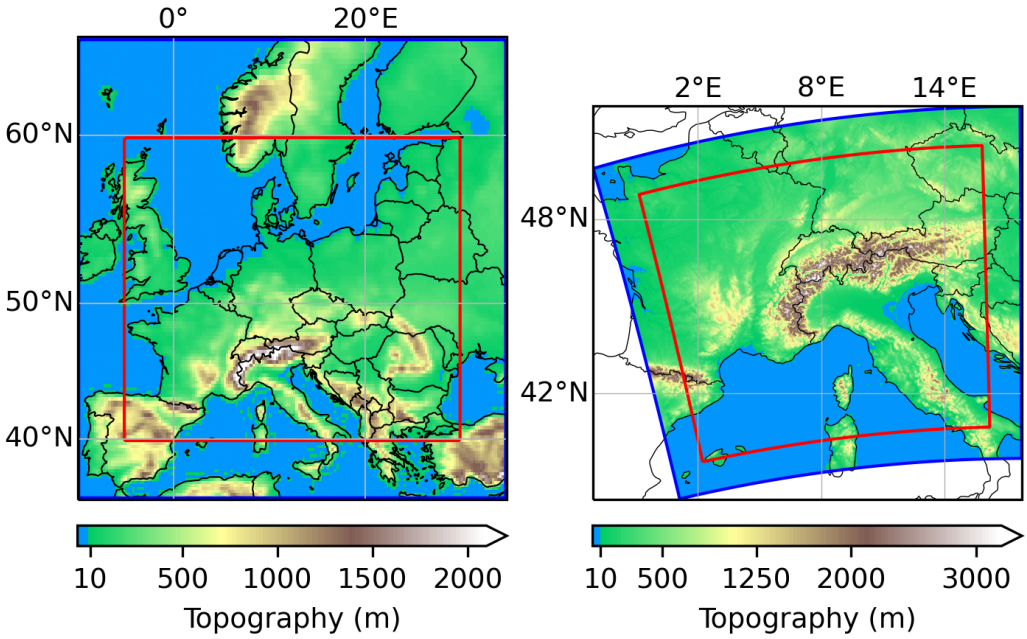

Figure S1: Topography of the considered regions for the ERA5 data (left) and the CP data (right). The blue and red boxes represent the considered input and target regions, respectively.

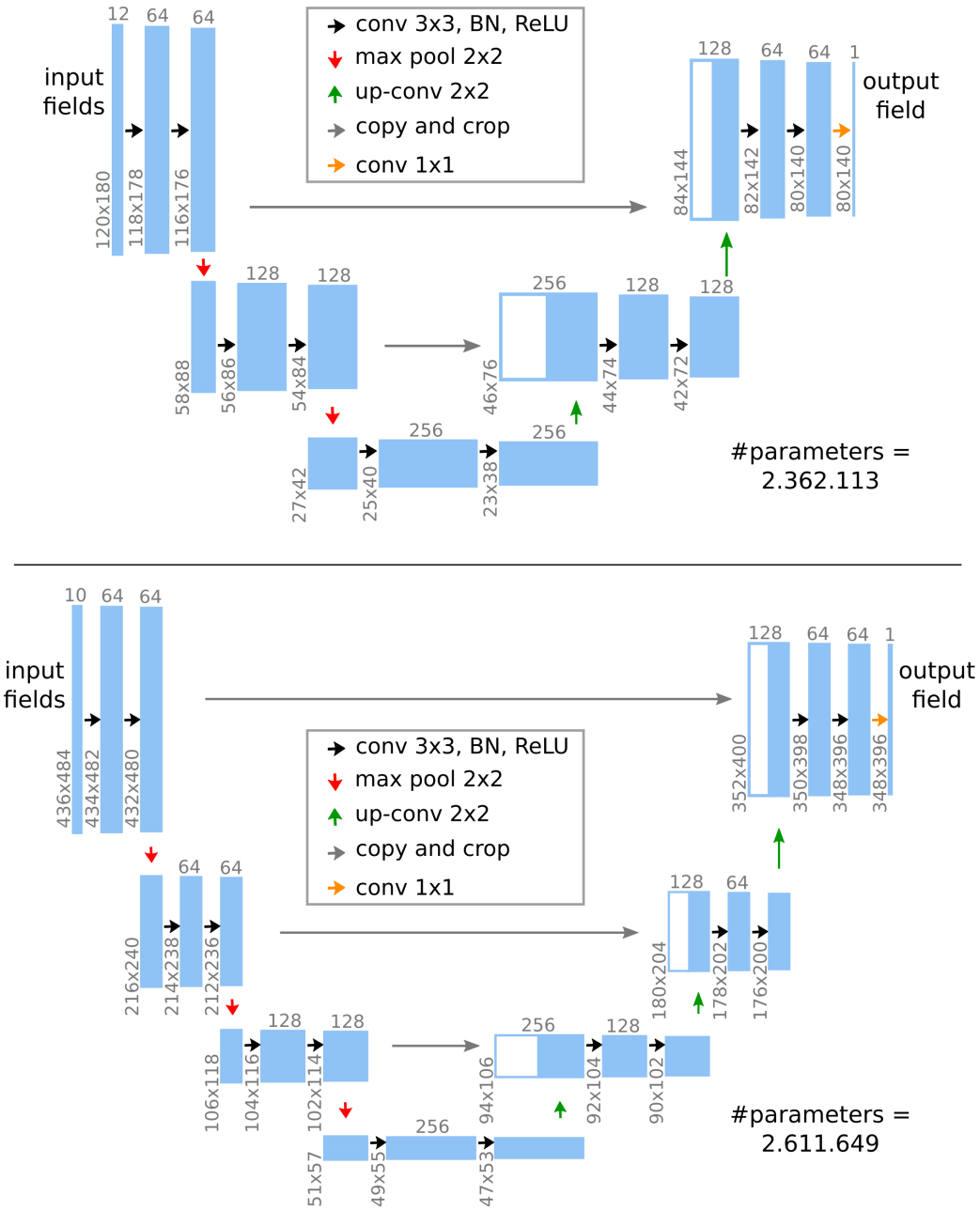

Figure S2: Model architectures considered for the ERA5 data (upper panel) and the CP data (lower panel) in the first experiment. The model architectures were inspired by the U-Net architecture (Ronneberger et al., 2015). The inputs to the models are represented by the leftmost blue boxes. They consist of 12 variables at the 120×180 input pixels for the ERA5 data, and 10 variables at the 436×484 input pixels for the CP data, respectively. The inputs are passed through multiple sequential modules, each of which performs simple mathematical operations on its respective inputs and produces an output that is fed to the next module as indicated by the arrows. In general, this output 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 boxes represent the outputs of the models, which consist of the precipitation predictions at the 80×140 target pixels for the ERA5 data, and at the 348×396 target pixels for the CP data, respectively. The only difference in the second experiment is the number of input variables (17 and 14, respectively).

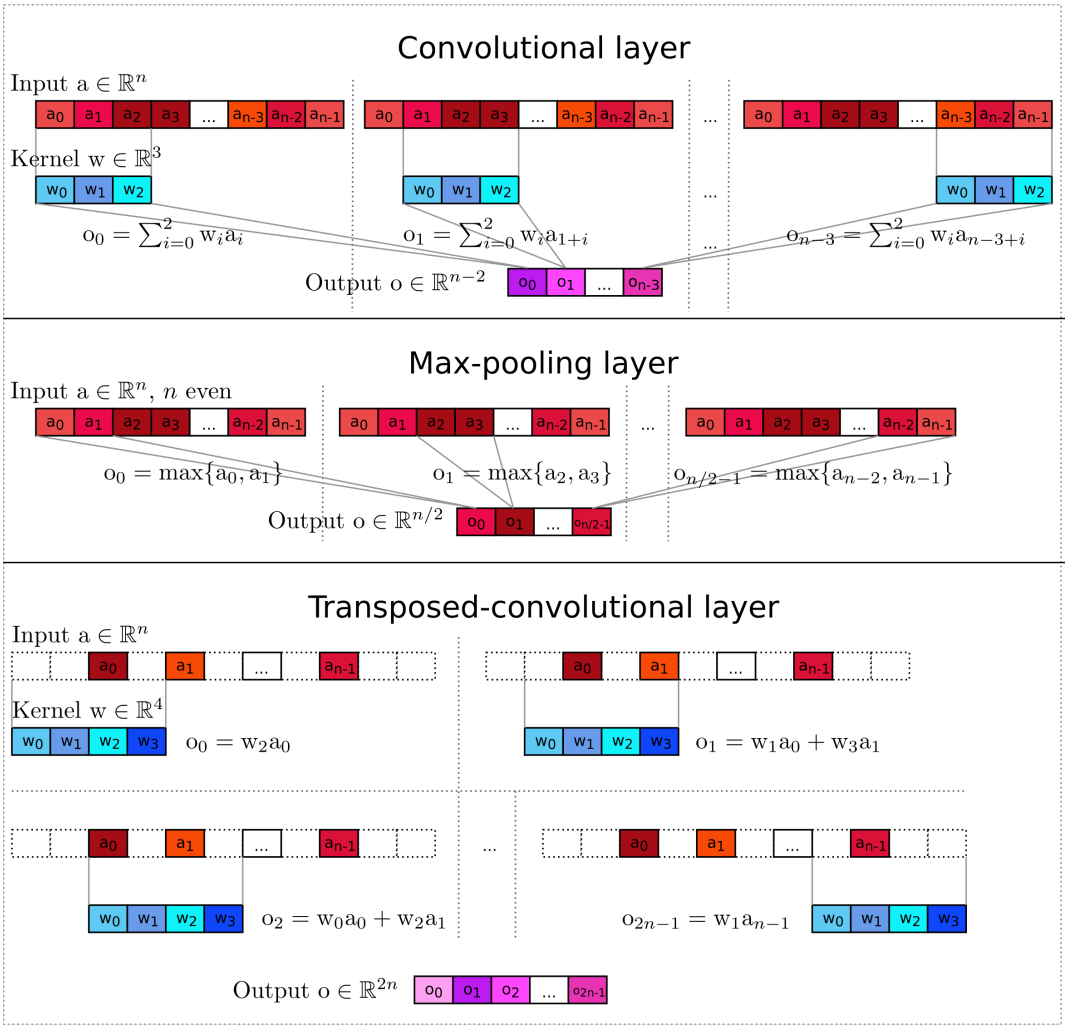

Figure S3: Model layers that affect the receptive field of a target pixel. The considered DL model (see Figure S2) contains three types of layers that affect the receptive field of a target pixel, namely convolutional layers, max-pooling layers and transposed-convolutional layers. Here, the effect of these layers on the receptive field is illustrated in the 1-dimensional analogue of the configurations used in this work. See section 2.2 and Text S2.

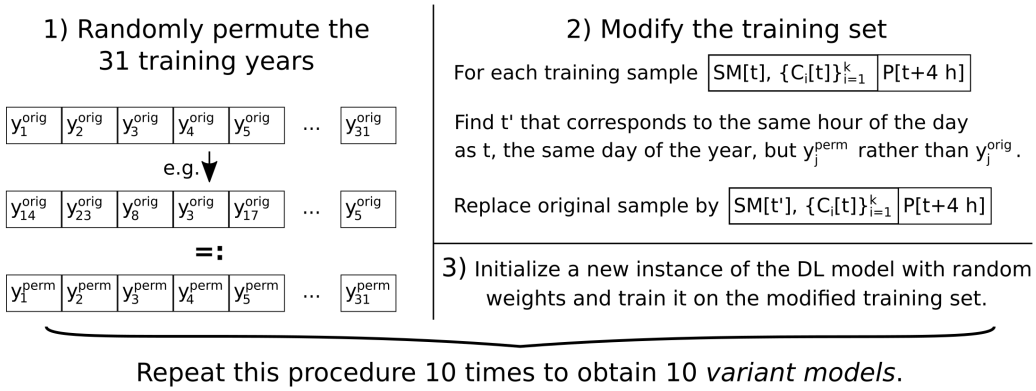

Figure S4: Procedure to obtain variant models for the significance analysis. See Text S3.

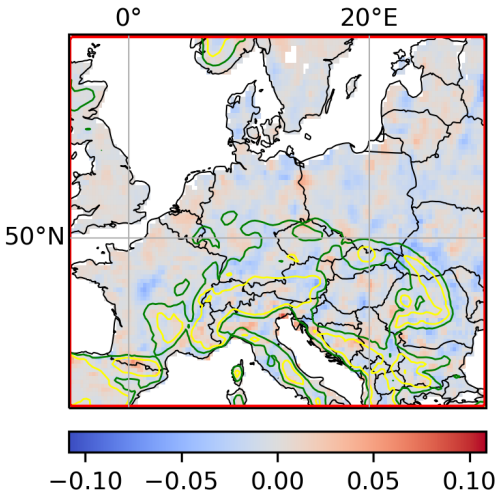

Figure S5: Pixel-wise difference in MSE on the test set between original and variant models for the ERA5 data in the first experiment. The difference is smaller than 0 for 60 % of all target pixels. See Text S3.

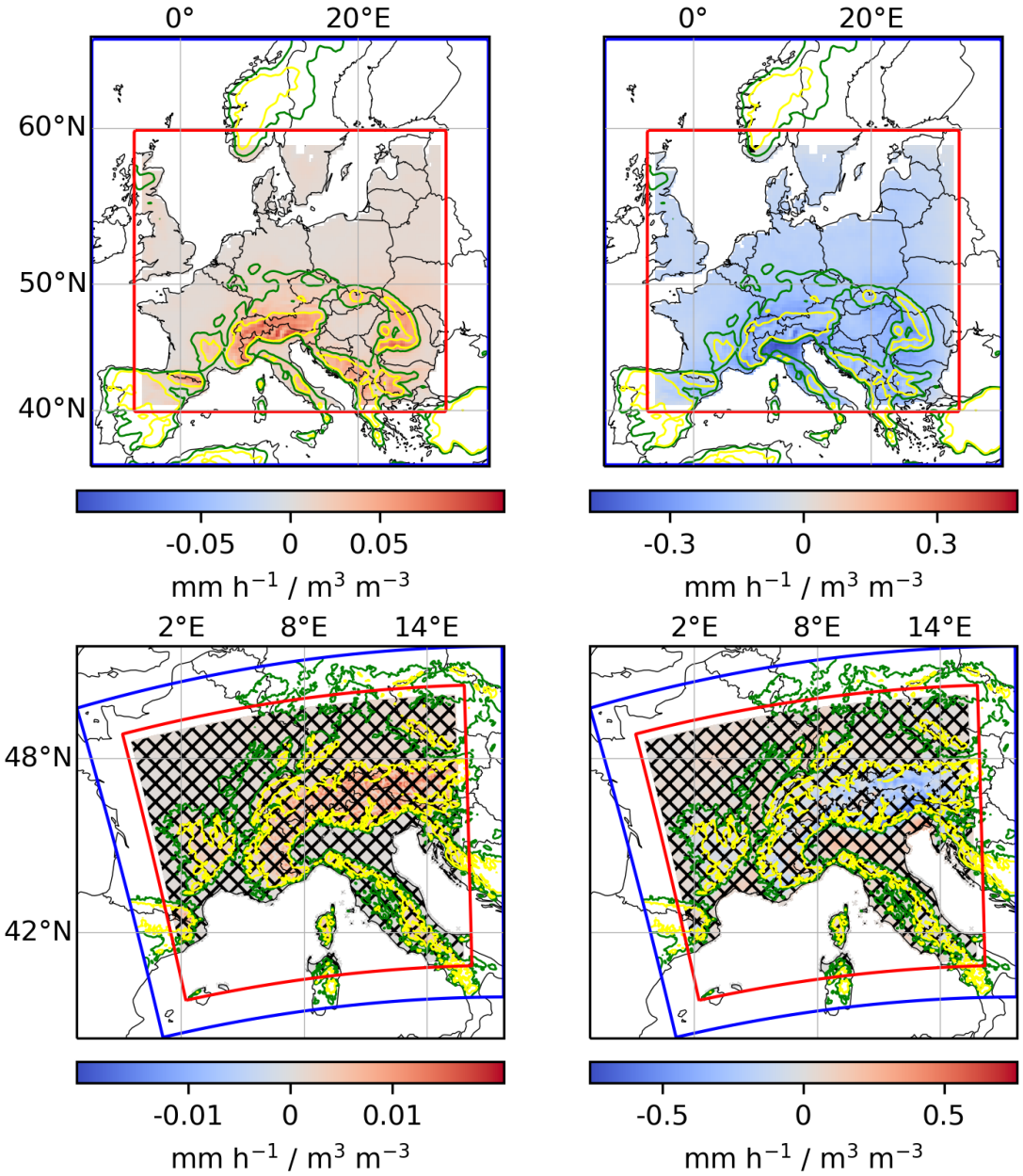

Figure S6: Local (left) and regional (right) SM–P couplings for ERA5 data (top) and CP data (bottom). Same as Figure 3, but for the second experiment, where we use more additional input variables (see Figure 2). For the ERA5 data, the confidence that the original models perform better than the variant models decreases from 1 in the first experiment to 0.84 in the second experiment. For the CP data, the confidence decreases from 0.9 to 0.74. We assume that the decrease is because the mapping of input variables to the expected value of the target variable is harder to learn for more input variables (Tesch et al., 2023, Sect. 3.3), and because of correlations between the additional input variables in the second experiment and SM, which may provide some information on real SM values to the variant models.

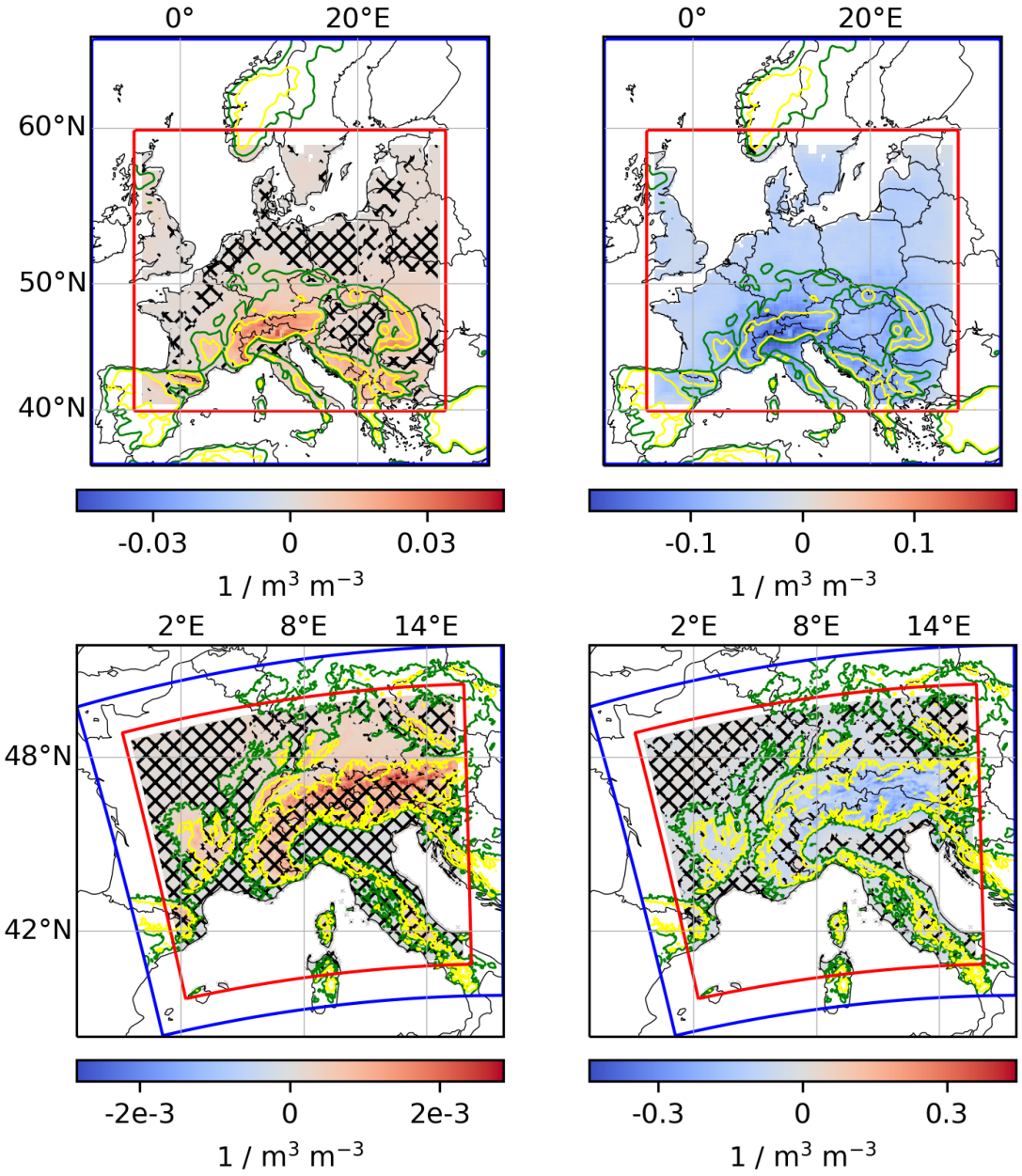

Figure S7: Local (left) and regional (right) SM–precipitation-*probability* couplings for ERA5 data (top) and CP data (bottom). Same as Figure 3, but for precipitation probability, i.e. for replacing the target variable $P[t + 4\,\mathrm{h}]$ by $P[t + 4\,\mathrm{h}] \geq 1$ mm (see last paragraph in section 3). Note that regional SM–precipitation-probability is defined analogously to regional SM–P coupling in section 2 (by building a sum of precipitation probabilities over single pixels). As for regional SM–P coupling, this was done for better comparability of the strength of local and regional couplings. However, it complicates the interpretation of regional SM–precipitation-probability coupling itself. For the ERA5 data, the original models perform better than the variant models with a confidence of 0.93. For the CP data, the confidence is 0.98.

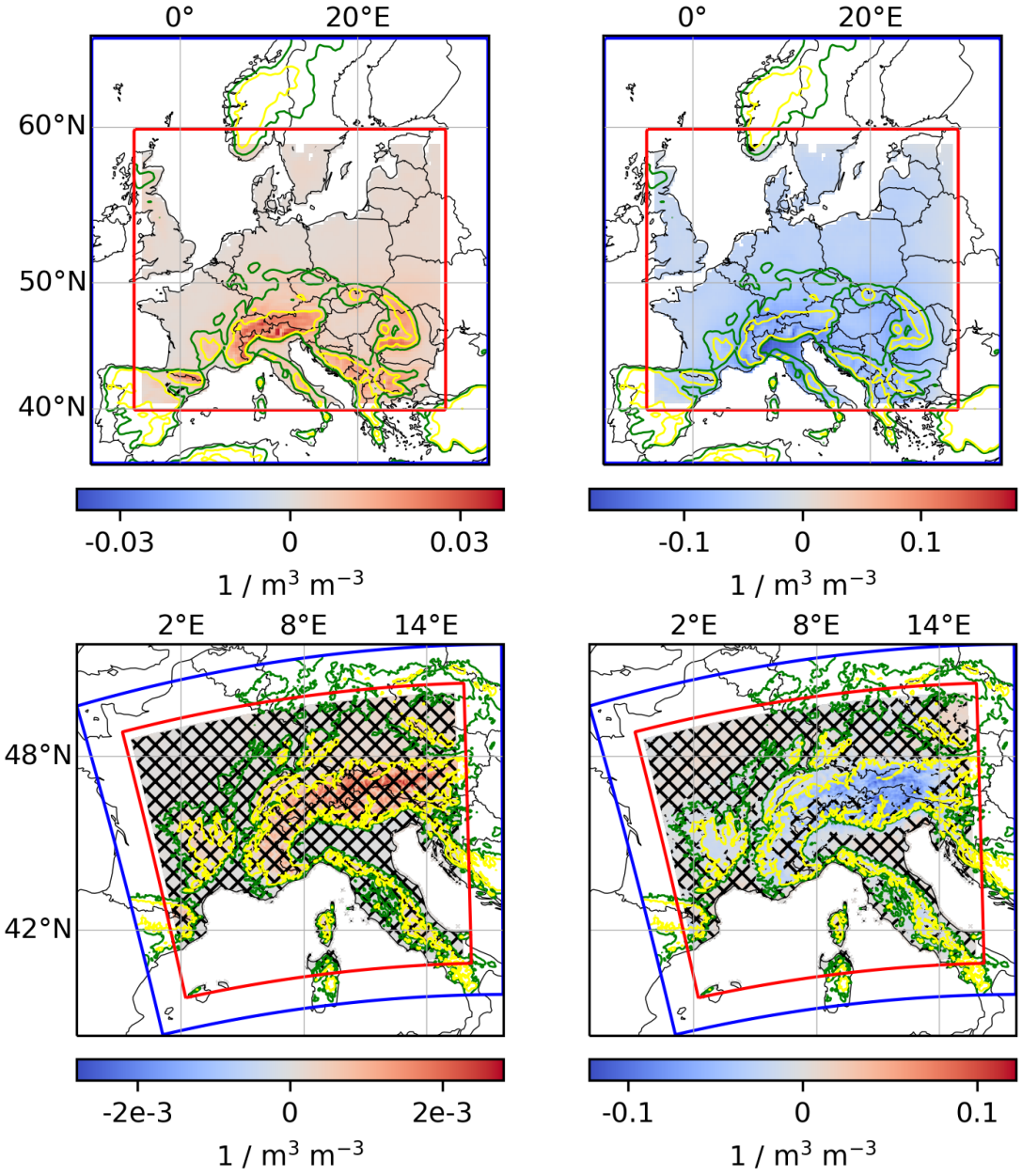

Figure S8: Local (left) and regional (right) SM–precipitation-*probability* couplings for ERA5 data (top) and CP data (bottom). Same as Figure 7, but for the second experiment, where we use more additional input variables (see Figure 2). The confidence that the original models perform better than the variant models is 0.17 for the ERA5 data and 0.08 for the CP data, which corresponds to a confidence of 0.83 and 0.92, respectively, that it is *larger* for the original models than for the variant models. This could indicate that, for some reason, the original models learned a statistical association between SM and precipitation that is not causal and does not generalize to the test set.

Table S1: Receptive field of a target index i for CP data. The receptive field of a target index i are the indexes in the input that could theoretically affect predictions at target index i. Last column: index that corresponds to target index i in the output obtained when only feeding the input variables in the receptive field to the considered DL model. The first seven and the last eleven target indexes are subject to boundary effects and neglected. See Text S2.

Case	Receptive field of target index i	Receptive field size	Index in small output corresponding to i
$i \bmod 8 = 0$	$i-8, \ldots, i+99$	108	8
$i \bmod 8 = 1$	$i-9, \ldots, i+98$	108	9
$i \bmod 8 = 2$	$i-10, \ldots, i+97$	108	10
$i \bmod 8 = 3$	$i-11, \ldots, i+96$	108	11
$i \bmod 8 = 4$	$i-12, \ldots, i+95$	108	12
$i \bmod 8 = 5$	$i-13, \ldots, i+102$	116	13
$i \bmod 8 = 6$	$i-14, \ldots, i+101$	116	14
$i \bmod 8 = 7$	$i-7, \ldots, i+100$	108	7

Table S2: Receptive field of a target index i for ERA5 data. As Table S1, but for the smaller DL model considered for the ERA5 data. For this model, the first and the last three target indexes are subject to boundary effects and neglected.

Case	Receptive field of target index i	Receptive field size	Index in small output corresponding to i
$i \bmod 4 = 0$	$i-4, \ldots, i+43$	48	4
$i \bmod 4 = 1$	$i-5, \ldots, i+46$	52	5
$i \bmod 4 = 2$	$i-6, \ldots, i+45$	52	6
$i \bmod 4 = 3$	$i-3, \ldots, i+44$	48	3

References

T. Hesterberg. What Teachers Should Know about the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. 2014. doi: 10.48550/arXiv.1411.5279.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. 2017. doi: 10.48550/arXiv.1412.6980.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc., 2019. doi: 10.48550/arXiv.1912.01703.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. doi: 10.48550/arXiv.1505.04597.

T. Tesch, S. Kollet, and J. Garcke. Causal deep learning models for studying the Earth system. *Geoscientific Model Development*, 16(8):2149–2166, 2023. doi: 10.5194/gmd-16-2149-2023.

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

List of Figures

1.1. The water cycle . . . 3
1.2. Methodological contributions in this thesis . . . 5
2.1. Schematic representation of a deep learning model . . . 9
2.2. U-net model architecture . . . 12
3.1. Areas and methodologies of causality research . . . 16
3.2. Structural causal model . . . 17
4.1. Concurring effects of soil moisture increases on subsequent precipitation . . . 24
8.1. Errors in the methodology of causal deep learning models . . . 50
A.1. Set up of the two original prediction tasks . . . 80
A.2. Location variant tasks . . . 83
A.3. Results for CNN1 in the rain prediction task . . . 85
A.4. Results for CNN1 in the location variant tasks . . . 86
A.5. Results for CNN2 in the rain prediction task . . . 87
A.6. Results for logistic regression in the rain prediction task . . . 87
A.7. Results for MLP in the water level prediction task . . . 88
A.8. Results for MLP in the location variant tasks . . . 88
A.9. Results for linear regression in the water level prediction task . . . 89
A.10. Results for linear regression in the location variant tasks . . . 89
B.1. Example for a causal graph . . . 101
B.2. Concurring pathways of soil-moisture—precipitation coupling . . . 104
B.3. Input and target regions in the example of soil-moisture—precipitation coupling . . . 104
B.4. Model architecture in the example of soil-moisture—precipitation coupling . . . 105
B.5. Causal graph in the example of soil-moisture—precipitation coupling . . . 106
B.6. Local and regional soil-moisture—precipitation couplings . . . 107
B.7. Linear correlation coefficient of local soil moisture and regional precipitation . . . 107
B.8. Modification of the training data for the year-wise permutation of SM[t] . . . 110
B.9. Soil-moisture—precipitation couplings for models trained on different training years 111
B.10. Soil-moisture–precipitation couplings for models trained on different months . . . 112
B.11. Soil-moisture–precipitation couplings for models trained for different regions . . . 112
B.12. Different regions considered . . . 112
B.13. Sum of soil-moisture–convective-precipitation and soil-moisture–large-scale-precipitation couplings . . . 113
B.14. Soil-moisture–convective-precipitation and soil-moisture–large-scale-precipitation couplings . . . 113
B.15. Product of soil-moisture–evaporation and evaporation–precipitation couplings . . . 113

B.16. Soil-moisture–evaporation and evaporation–precipitation couplings 113
C.1. Concurring effects of soil moisture increases on subsequent precipitation 123
C.2. Causal graph summarizing the choice of input variables 127
C.3. Local and regional soil-moisture–precipitation couplings 129
C.4. Topography of the considered regions 146
C.5. Model architectures 147
C.6. Model layers affecting the receptive field of a target pixel 148
C.7. Procedure to obtain variant models for the significance analysis 148
C.8. Pixel-wise difference in MSE between original and variant models 149
C.9. Local and regional soil-moisture–precipitation couplings in the second experiment 150
C.10. Local and regional soil-moisture–precipitation-probability couplings 151
C.11. Local and regional soil-moisture–precipitation-probability couplings in the second experiment 152