universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Exploring Classifier Skill in Distinguishing Climate Model and Observational Data

verfasst von | submitted by

Julian Merio BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 910
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Computational Science
Betreut von | Supervisor:	Dr. Lukas Brunner BSc MSc

Abstract

Climate models are important tools for gaining insights into the processes that shape the earth's climate. These physical models, used for conducting experiments that run on supercomputers, are constantly improved, providing more realistic simulations from generation to generation. However, climate models are developed around the world and exhibit differences in the plausability of their outputs.

For improving these climate models, climate model evaluation plays a fundamental role. It is a complex field dealing with critical questions such as data uncertainty. The basic approach in climate model evaluation is to compare the climate model output to actual observations. Regarding the problem of observational uncertainty, it becomes far from trivial to define thresholds at which one would reject a climate model. Moreover, there are numerous varibales that need to be considered. Climate models might simulate parts of the meteorological process very accurately while performing poorly on others. Hence, a given climate model can also be used for simulating only parts of the climate dynamics, where the model is known to be accurate.

The application of machine learning methods is becoming increasingly popular in general and has also been successfully used in climate science. Brunner and Sippel (2023) have shown that a Convolutional Neural Network (CNN) is capable of robustly identifying observational and climate model data using daily temperature maps. Even after removing bias by subtracting the mean seasonal cycle from the training data, the CNN can accurately make the distinction at such short timescales. A drawback of the classical evaluation methods is that the average over several decades must be computed for the climate model output in order to reduce internal variability and obtain comparable tendencies. This, in return, leads to a demand for more observational data. Machine Learning methods could overcome this issue, since they can effectively learn dataset specific differences from such short timescales as shown by Brunner and Sippel (2023). This raised the question, what patterns are actually learned and what geographical regions might be the most important ones for a skillful prediction. In particular, it was worth investigating how the learned patterns change upon the removal of the mean seasonal cycle.

To gain a deeper understanding of the decisions made by the CNN, explainable artificial intelligence (XAI) was employed. Through the application of layerwise relevance propagation (LRP), an algorithm that propagates backwards from the output layer to the input layer of the CNN, it was possible to obtain explanations showing the geographical regions that are most relevant for the CNNs' decision. Regions such as e.g. the north Atlantic, the equatorial Pacific or the southern Ocean emerged as important regions during the experiments. The resulting explanations were evaluated by making use of an input perturbation method, showing the reliability of the results. Further investigations of the misclassifications of the deep CNN classifier revealed the influence of the training data on the prediction accuracy. Moreover, it could be shown that the explanations show strong dataset specific differences when using various observational datasets. In general, the effect of removing the mean seasonal cycle clearly affected the explanations obtained by LRP. However, it could be found that some regions are highly relevant even after removal of the mean seasonal cycle, while the distribution of the important input features or even their type of contribution to the CNN classifiers' decision (positive or negative) changed.

The results contribute to a better understanding of XAI in the context of climate science. These findings demonstrate the complexity of analyzing the decision-making process of a CNN or artificial neural network in general. The application of these techniques indeed provide an excellent first step towards the understanding of the CNNs' decisions. Additional insights would require a much deeper analysis, especially for highly complex tasks such as climate model evaluation.

Kurzfassung

Klimamodelle sind wichtige Instrumente, um entscheidende Einblicke in klimatische Prozesse zu gewinnen. Diese physikalischen Modelle stellen aufwändige Berechnungen dar. Sie werden verwendet, um Simulationen auf sehr leistungsstarken Rechnern auszuführen. Diese immer besser werdenden Modelle werden auf verschiedensten Instituten weltweit entwickelt und unterscheiden sich somit auch in der Plausibilität ihrer Ergebnisse.

Für die Verbesserung der Klimamodelle spielt die Evaluierung eine fundamentale Rolle. Dabei handelt es sich um eine komplexe Aufgabe, für die es wichtige Probleme zu lösen gilt, nicht zuletzt die Bestimmung der Unsicherheiten in den Daten. Die grundlegende Herangehensweise sieht vor, die Ergebnisse von Klimamodellen mit realen Beobachtungen aus der Vergangenheit zu vergleichen. Probleme wie die Unsicherheiten in den Beobachtungsdaten führen dazu, dass es genauere statistische Analysen erfordert, um Grenzen festzulegen, ab denen ein Modell als exakt gilt oder nicht. Außerdem gibt es zahlreiche Variablen die in die Simulationen bzw. Berechnungen miteinbezogen werden müssen, was zu einer hohen Komplexität führt. Dadurch kann es sein, dass bestimmte Klimamodelle nur für einzelne physikalische´Prozesse genutzt werden, da sie nur für einen Teil der Simulationen exakte Ergebnisse liefern.

Machine Learning (ML) Methoden finden immer häufiger Verwendung in den Klimawissenschaften. Brunner and Sippel (2023) haben gezeigt, dass ein Konvolutionales Neuronales Netzwerk (CNN - Convolutional Neural Network) in der Lage ist, mit hoher Genauigkeit zu unterscheiden, ob ein Ergebnis von Klimamodellen oder Beobachtungsdatensätzen stammt unter der Verwendung von globalen Tagestemperaturkarten. Dieses CNN war auch nach Entfernung wichtiger zugrunde liegender Verzerrungen in den Trainingsdaten in der Lage, diese Unterscheidung genau zu tätigen. Diese Entdeckungen warfen die Fragen auf, welche zugrundeliegenden Muster so ein CNN von den Daten lernen kann und zudem, welche geografischen Regionen dafür am ausschlaggebendsten sind.

Um ein tiefgründigeres Verständnis für die Entscheidungen des CNNs zu bekommen, wurden im Zuge dieses Projektes Methoden der erklärbaren künstlichen Intelligenz (XAI - explainable artificial intelligence) eingesetzt. Genauer wurde die Methode Layerwise Relevance Propagation (LRP - Schichtweise Relevanzpropagation) verwendet, ein Algorithmus, der ausgehend von den Ergebnissen die das CNN vorhersagt, rückwärts durch die Schichten des Neuronalen Netzes propagiert und die einzelnen Gewichtungen der Merkmale zurückverfolgt. Das Ergebnis zeigt am Ende, welche Merkmale für die exakten Vorhersagen des CNNs besonders wichtig sind. Zu diesen zählen u.a. der Nordatlantik, der Äquatorialpazifik oder der südliche Ozean. Damit war es möglich, die geografischen Regionen auszumachen, die das CNN besonders benötigt, um die Klimamodelldaten von den Beobachtungsdaten zu unterscheiden. Diese Resultate, die 'Erklärungen' der Vorhersagen, wurden weiters mithilfe von Perturbationsmethoden evaluiert. Diese haben gezeigt, dass die Ergebnisse tatsächlich entscheidend für eine genaue Unterscheidung bzw. Vorhersage sind. Zusätzliche Untersuchungen der falschen Vorhersagen haben gezeigt, wie die Genauigkeit zudem von den spezifischen Trainingsdaten abhängt. Außerdem konnte gezeigt werden, dass diese Erklärungen auch schon im Falle eines binären CNNs Unterschiede innerhalb der einzelnen Datensätze (Modelle u Beobachtungen) aufweisen können. Dies galt vor allem für die Beobachtungsdatensätze im Zuge dieses Projekts. Generell

war es augenscheinlich, dass die Entfernung von wichtigen Verzerrungen, wie in diesem Fall durch das Entfernen des mittleren saisonalen Zyklus, das Lernverhalten des CNNs erheblich beeinflusst. Die räumliche Verteilung der wichtigen Merkmale hat sich teils stark unterschieden.

Die Ergebnisse dieser Arbeit tragen zu einem besseren Verständnis von XAI Methoden im Kontext der Klimawissenschaften bei und demonstrieren wie komplex sich die Analyse des Entscheidungsprozesses von einem CNN gestaltet. XAI Methoden stellen einen wichtigen ersten Schritt für das Verständnis komplexer ML Methoden dar, um deren Entscheidungen besser nachvollziehen zu können. Dennoch benötigt es für komplexe Probleme wie die Evaluierung von Klimamodellen weitere Schritte, um die Resultate dieser XAI Methoden tiefgründiger zu verstehen.

Acknowledgements

A big thanks goes out to my supervisor Lukas Brunner, who proposed this topic to me and always took the time to answer all my questions. Being available for any concerns while giving me enough freedom for experimentation made up the perfect supervision for me. It was an exciting project that taught me a lot about climate data.

Thanks to Sarah for always giving me honest feedback.

Contents

1 Introduction

The following introduction shall provide the basic knowledge necessary to capture the essence of this project. It shall give an understanding of climate models at the beginning, explaining the importance of climate model evaluation as well and roughly how this is done. The application of machine learning in climate science is shortly touched, bridging the gap to the fundamentals of machine learning in general and the basic concepts behind Artificial Neural Networks. The goal is to make the findings accessible to the different fields involved, since this projects can be located at the intersection between Data Science and Meteorology.

1.1 Climate Models: Function and Importance

Climate models are essential tools for scientists to understand the Earth's climate. These are very diverse and cover different parts of the climate system. They are based on physical equations which form the core of a climate model. By translating these physical principles into numerical methods, the aim is to approximate good enough solutions. Since these equations include complex partial differential equations as well, such as the Navier-Stokes equations (describing viscous fluid dynamics), which still lack unique solutions, models can only approximate (CarbonBrief 2018). Consequently, climate models require significant computational power for their simulations. Today's complex climate models are huge programs solving all the involved physical equations on supercomputers. Climate Models are developed around the world and slightly differ in the implementations of these approximations, leading to different results. Also, there are very small-scale processes which cannot easily be translated into equations, which are handled differently ("Parameterization").
The basic approach is to divide the digital version of the earth into grid cells, which function as units for computation. For each of these grid cells, the approximate solutions of the mentioned physical equations are computed. These equations are grounded in fundamental principles of physics, such as classical mechanics, thermodynamics, and electromagnetic radiation (Gettelman and Rood 2016). Climate models do not operate autonomously; they require initial conditions known as "forcings" to begin simulations. Forcings, such as CO2, methane, and NO2, influence the energy absorbed by the Earth and retained in the atmosphere and are usually based on estimations. These greenhouse gases, where water vapor is the most abundant greenhouse gas, heavily influence the energy flows within the energy budget of the atmosphere and therefore the climate system in general. A precise estimation is therefore of high importance to reflect reality as closely as possible within a climate model. Most modern models use "Representative Concentration Pathways" (RCPs), which provide plausible future scenarios based on socio-economic developments (CarbonBrief 2018). Those estimations are crucial for running a climate model since they represent the initial conditions from which a model starts its simulations.
Regarding the output of climate models, one can have results for various physical climate variables, since there are climate models out there simulating all the different components of our climate system. Since our climate system is best seen as the distribution of weather states (Gettelman and Rood 2016), knowing the variability within this distribution is important for making any assumptions based on climate model simulations. Moreover, by holding certain forcings as e.g. the greenhouse gases at a constant concentration, it

is possible to estimate the climate sensitivity. This means that it can be evaluated how strong the system reacts to certain concentration levels or alterations. Running such experiments with climate models, scientists can obtain estimates about the impact of certain changes in our climate system, allowing for future projections, which can then be used for estimating the effect of a given scenario.

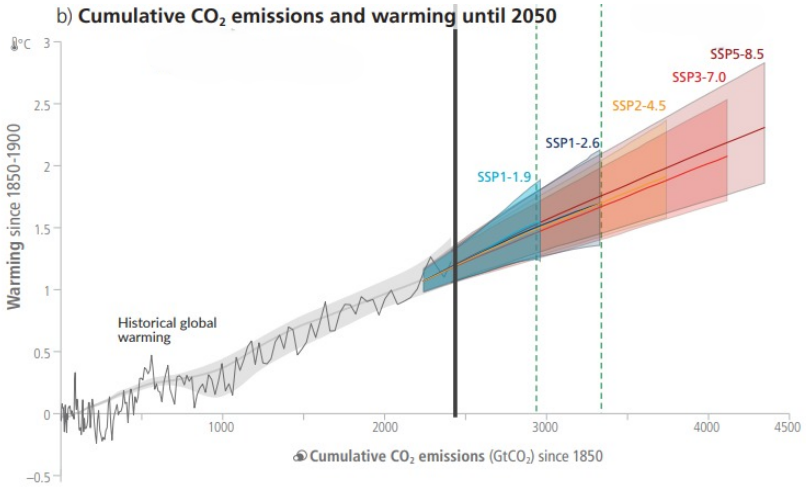


Figure 1: *Taken from the IPCC AR6 Synthesis Report (2023). Example of a typical plot combining historical data and climate model projections. The figure shows the cumulative CO2 emissions vs. the surface temperature increase under the different Shared-Socioeconomic-Pathways (SSPs), starting from 1850. The grey part represents the historical CO2 emissions vs the observed temperature data until 1900. The shaded areas show the most likely range for the temperature. The colored parts show the projections for the different CO2 emission scenarios (SSPs). The vertical black line represents the year 2020 while the dashed lines mark the carbon budgets for a temperature increase of 1.5 and 2 degrees celsius.*

1.2 CMIP and Climate Model Evaluation

Evaluating climate models is essential for improving them and hence obtain more realistic simulations of the earth's climate - but how is that done?

Evaluation of climate model output has become an increasingly complex task due to the rapid growth in the number of climate models belonging to the 'Coupled Model Intercomparison Project' (CMIP). CMIP started off in 1995 with the idea to make all the different climate model simulations from different modeling centers comparable. Since every simulation is set up differently as well as the model itself, the differences in the output can be due to both the experimental set up or the models themselves. Therefore CMIP experiments are conducted to gain insights into the model differences and commonalities. For example, the models are run with the same initial conditions to verify whether differences in the output originate from differences between the models and not the experimental set up. Every 5-6 years a new CMIP generation appears with increasingly sophisticated experiments conducted (CarbonBrief 2018). The goal of this project is also to introduce standardization where possible between all the existing climate models which have to fulfill specific requirements. The project aims to collect the models and the corresponding experiments, each model center runs. This has the great benefit that

for any question one might have, all the results from various experiments can be used for comparing the climate model outputs, which makes estimations of the variability or model uncertainty possible. Since CMIP is seen as an "ensemble of opportunities", meaning that the project follows an inclusive approach including all the models meeting the simulation guidelines and requirements, the number of models participating in CMIP has grown from generation to generation (Merrifield et al. 2023). This fact in particular makes the evaluation of climate models more important since the models differ in their setup for the simulations. While some models are only represented by a unique simulation, others try to sample the model uncertainty by simulating several times, also changing the initial conditions and using different models for the same task (representative sample of model uncertainty). Some modeling centers also provide several versions of the same base model with e.g. different resolution levels. These are some of the crucial facts which highlight the importance of climate model evaluation, since these differences ultimately lead to different climate model output. Moreover, through the constant improvement of the models, more realistic simulations have been obtained, but it has still been reported that not all of them are equally credible (Annan and Hargreaves (2017); Eyring et al. (2019)). Models which are part of CMIP are an important source for future projections of our climate system and are featured in today's reports of the IPCC.

The basic approach for evaluating climate models is to compare them to observations (Gettelman and Rood 2016). A common approach for evaluating climate models is the so-called hind-casting. This involves the process of letting a climate model predict observed periods in the past. This shows scientists whether the model correctly predicts the already observed conditions, which makes accuracy estimates possible. This approach requires large amounts of data since the classical procedure is based on multi-decadal averages. This is done to average out the internal variability (i.e. weather) to be able to compare the climate models' ability of simulating the climate to the observations. Evaluations mostly have some aspect of the climate model in mind, since it is not possible to realistically evaluate everything a model simulates at once. The Global Circulation Models or Global Climate Models (GCMs) which are part of CMIP simulate the whole atmosphere amongst others, involving many complex processes. These models are the results of many smaller models coupled together, each one computing different aspects of the climate dynamics. For that reason, mostly one variable is taken to evaluate against the observations, which is also done in this work, as the temperature maps of the models are used (see 3.1).

Regarding the observations, it has to be stated that they also involve statistical operations and contain errors (Zumwald et al. 2019). Therefore, we have to deal with observational uncertainty as well and they can be seen as little models to different extents - their underlying principles also cover different approaches, depending on the type: from reanalysis datasets (involving numerical weather predictions) to more direct observational products which e.g. just apply some interpolation on the observations measured as in e.g. the DOISST dataset. Reanalysis datasets involve data assimilation by using observations from a variety of sources and combining them, as well as numerical weather prediction (NWP) (Parker 2015). NWP makes a first guess which is subsequently updated by the real world observations. Reanalysis datasets therefore are obtained by data assimilation for past periods, covering large time spans, making snapshots with respect to sub daily intervals. So the classical approach for model evaluation involves several steps. A crucial part is to understand the uncertainties within the observations. Since the climate model output is compared to some sort of observational dataset, one has to define a decision

threshold for determining whether a climate model is accurate or not. As long as the differences between the model and the observation lies within these limits of variation defined by the uncertainty (+/- some degrees regarding temperatures for instance) the model can be considered as accurate. Therefore, the observations being used have to be analyzed before any evaluation procedure. Regarding reanalysis datasets for instance, there are many potential sources of uncertainty which can add up. The instruments used for measuring and the NWPs all contain errors, which are hard to understand and consider within the statistics. A reasonably accurate estimation of the uncertainty within the observations is nevertheless crucial for the evaluation. An important note at this point is therefore, that the mean state of the model and observational outputs alone is not sufficient, the distribution of climate states need to be right. This means that the variance within the models has to match those of the observations as well as possible (Gettelman and Rood 2016). In climate model evaluation though, decadal averages are often used to eliminate the effect of random weather, but this approach doesn't take care of the variability among the datasets.

1.3 Machine Learning speeding up Climate Model evaluation

Knowing all the necessary steps for climate model evaluation, it might become clear that it is a statistically very sophisticated task involving many uncertainties and important decisions. It is definitely not trivial to study the uncertainties in all the datasets, especially the observations. And still, by only looking at one parameter as the temperature for example, there might be underlying biases, which are hard to discover. Spatial patterns are hard to disclose by using rather classical statistics. A drawback of the classical evaluation methods is that the average over several decades has to be computed for the climate model output in order to reduce internal variability and obtain comparable tendencies. This, in return, leads to a demand for more observational data. Also the hard work of trying to understand the uncertainties within the observations is quite time consuming. Recently, the methods of machine learning have already been proven to successfully extract patterns among climate model and observational data already at much shorter timescales (Brunner and Sippel (2023), Sippel, Székely, and Knutti (2020)). The work by Brunner and Sippel (2023) has shown that a CNN can identify climate model or observational data in a robust manner, when trained on daily temperature maps. This would mean that these CNN's can already learn strong underlying patterns at such short timescales. Given that data can be directly passed to a CNN without relying on decadal averages or additional statistical measures, these networks could accelerate the fundamental process of comparing climate model outputs with observational datasets. Furthermore, CNNs have the potential to provide deeper insights into the various datasets as a whole. Nevertheless, without knowing the rough patterns such a CNN learns, it will be hard to make any conclusions for improving climate models. Therefore, it is important to gain insights into the important features that play a crucial role or ideally the learning process in general. The task of getting information about the learning behaviour of such a CNN belongs to a subfield of AI called explainable AI. This will be discussed later in more detail. However, it shall be mentioned that XAI methods would play a decisive role for understanding the differences between climate model output and observational data. These methods would ideally provide explanations of the decisions a CNN makes. Assuming that a given CNN such as the one by Brunner and Sippel (2023) is very accurate, the explanations could lead us towards important tendencies among the various datasets

used in climate science.

Since more and more data is produced which can be used, the demand of machine learning applications increases as well as the number of successful use cases. Among the most successful machine learning methods, the CNNs take a very prominent position. Their capability of capturing high dimensional, non-linear relationships in data has often been reported, besides the successful beginnings in image recognition tasks.
The following part has the aim to give the reader a good intuition how such Artificial Neural Networks (ANN's) roughly work and tries to make clear what components are important for the architecture of Convolutional Neural Networks (CNNs). Additionally, in order to define the term machine learning, a broad definition as well as a short background will be given.

1.4 Overview of Machine Learning

The beginning of Machine Learning even dates back to the 1940s (Goodfellow, Bengio, and Courville 2016). Therefore, it is everything but new; it just went through different phases of popularity and success. Generally speaking, to cite one of the pioneers in this field, Tom Mitchell (1997, as cited in Goodfellow, Bengio, and Courville (2016)) defines machine learning as "a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." This definition includes any computer program that improves by experience.
This definition still holds, as the current algorithms still follow this basic principle. The two fundamental paradigms on which almost all the machine learning approaches build on are called supervised and unsupervised learning (O'Shea and Nash 2015). Of course, there are also transitions between these two (e.g. semi-supervised) or alterations such as reinforcement learning, which is considered as another paradigm, but these two can be considered as the most basic categories.
Supervised learning comprises all the methods working with labeled data, meaning that it receives the data together with the correct label. Assuming one wants to have a machine learning classifier distinguishing between pictures of cats and dogs, it would mean that the classifier gets the pictures together with the label 'dog' or 'cat'. This is referred to as the training step.
During unsupervised learning, the algorithm learns without labels, looking for commonalities within the data. Many clustering algorithms can be assigned to the category of unsupervised learning for instance. During clustering the data is typically grouped together, often based on specific distance measures, so there is no need for labeling the data. Since further treatment of the fundamental principles of machine learning goes far beyond the scope of this work, the focus will now be on the method actually used here: Convolutional Neural Networks (CNNs).

1.5 Neural Network Fundamentals

CNNs are a special case of Artificial Neural Networks (ANNs). ANNs all share the same basic structures, but depending on the specific tasks, they undergo certain modifications. ANNs try to mimic the biological neurons to some extent, the architecture is influenced by the biological nervous system (O'Shea and Nash 2015). The basic unit, the neuron,

builds up the ANN. These neurons are organized into layers. In the most basic case there is one input layer, one hidden layer and one output layer. Depending on the tasks, the layers are designed differently. This mostly happens in the hidden layers, since they define how the information is being extracted from the data. These are called hidden, because the results of all the computations are not visible, while for the others this is the case (Goodfellow, Bengio, and Courville 2016). The input layer receives the data as it is while the output layer provides the result. Each node can be seen as its own linear operation, performing a regression task. This means every neuron gets some input from a previous neuron. If this input exceeds a previously determined threshold, the neuron is activated and computes the next output value and so on. Therefore, an ANN is a sequence of functions (represented by neurons), which are interconnected and organized into layers (Goodfellow, Bengio, and Courville 2016).
Another very important principle for the learning process of an ANN, is the computation of the error or loss function. Again, there is a pool of opportunities to choose from, but the important information here is that the error is constantly computed during the training process. Since the Network predicts some result for each sample during the training process, this allows it to check if the predicted result of the network is correct. Minimizing the loss function involves adjusting the weights of the neural network. For example, in image recognition tasks, each pixel of an input image is assigned a weight. These weights are modified during training, as some regions of the image are more important than others for distinguishing between categories like cats and dogs. The algorithm will, sooner or later, adapt the weights in a way, as accurate as possible. The weights are iteratively adapted in terms of the so-called loss function. The learning process heavily underlies the computation of gradients, which are computed by back-propagation. Back-propagation basically means that the chain-rule is efficiently applied to see how the changes in one node affect the outcome (Goodfellow, Bengio, and Courville 2016). Back-propagation is an important part of the optimization process, because it quantifies the level of influence of a certain change within a node with respect to the final output. This is a fundamental concept for the iterative updating of the weights of the Neural Network and can be seen as the heart of the learning process. The gradients computed via back propagation can be used for the optimization of the loss-function, minimizing the error of the predictions. For this, typically methods like the Stochastic Gradient Descent are used, to name just one popular optimizer.

1.6 Features of a CNN

CNNs were the ANNs which achieved the first important successes especially in image recognition. Their successful application can be considered as important incidents for the acceptance of neural networks in the broader scientific community and beyond (Goodfellow, Bengio, and Courville 2016).
As the name already suggests, the CNN deploys convolutional layers within its network. The other important component tailored to the demands of this type of network is the pooling layer (O'Shea and Nash 2015). A fully connected layer is also part of CNNs, but these are usually part of every ANN. In this case, every output of every neuron is passed to every neuron of the consecutive layer. In convolutional layers, not all of the neurons are connected. Without going too much into detail here, the basic functions of these layers shall be outlined. The convolutional operation within the corresponding layers involves a filter, which is often called the kernel. This kernel has a specific size and each

layer makes use of a number of kernels. These, ultimately being matrices gliding over the input, are able to capture spatial patterns in the form of activation maps. Therefore, the convolutional layers are responsible for the actual pattern recognition within the data. This makes them especially useful in image recognition, which is basically done within the scope of this work, since daily temeperature maps are used, where every grid cell can be considered a pixel.
The pooling layer summarizes those activation maps, by taking only the maximum value computed by a given convolutional layer around some area defined by the kernel size. This would be the case for max-pooling (Goodfellow, Bengio, and Courville 2016). This helps in reducing the complexity of the output. These two features, although including a bunch of non-trivial methods, are important for the architecture of a CNN. Still, there is a lot to consider when creating a CNN. But this is always dependent on the task and the case for any ANN.

1.7 Machine Learning in Climate Science

There is a variety of frameworks available which allow for a quick and relatively easy implementation of machine learning methods. Frameworks such as Tensorflow, Scikit-learn or Pytorch smoothed the way to a comparably easy handling of these methods. This is definitely a reason for its increasing use in many different scientific disciplines. Through existing code which can be quickly adapted to different scenarios, scientists can still focus on their domain while using these methods as a simple tool. Since CNNs are particularly useful in spatial pattern recognition, it is a popular choice when geographical data is involved.
As mentioned earlier (section 1.3), the application of machine learning in climate science has turned out to be useful for many tasks. As outlined in 1.2, the classical approach of climate model evaluation involves tasks which could be sped up. As shown by Brunner and Sippel (2023), a trained CNN is able to robustly distinguish between observational products and climate model output. This comparison, as mentioned earlier, is the basic approach in climate model evaluation. The data used were daily temperature maps, so the CNN was able to make a reliable distinction already on such short timescales compared to the windows of 20+ years commonly used in climate model evaluation (Brunner and Sippel 2023). So, once trained, a CNN is obviously able to extract differences between those datasets already from daily samples. This could help in accelerating the evaluation steps since important decisions as the 'decision-threshold' (outlined in 1.2) deciding whether an observation deviates enough from a model or not, don't have to be made. Still, it is important to state that such a CNN is not telling why and how the distinctions are made. Furthermore, the question of observational uncertainty is not solved at all by making use of such a CNN. But it can be used as a pointer showing that differences emerge already at daily samples, as in this case. This can help scientists to quickly get to important issues which need to be solved with respect to climate model performance. The application of a CNN would be a first step towards understanding what the differences may be in such a high dimensional setting. Further investigations of such CNNs and their decision making process have to be made as well as statistical operations in order to get really valuable insights. Assuming that we are once able to derive a precise reasoning for the CNNs decisions, this would probably facilitate a huge progress in the accuracy of climate models.

1.8 Explainable Artificial Intelligence

The increasing popularity of machine learning methods, particularly convolutional neural networks (CNNs), among climate scientists has raised the important issue of such an ANN being a black box. The fact that these ANNs are able to learn from massive data amounts, extracting information from high dimensional data, is obviously applicable to many different scenarios. But it is important to mention that the exact learning process, taking place in an iterative manner, is enclosed to the outside world. Technically speaking, for a CNN this means that we don't get any insights, how the weights of the specific input features are adjusted such that the CNN is able to make these predictions (the fundamental learning process by backpropagation 1.5). This issue is tackled by the discipline of eXplainable Artificial Intelligence (XAI). XAI has emerged as a subfield of AI, with the goal to build more reliable Neural Networks (Samek and Müller 2019). There are already different methods available, which makes it difficult to find out which best suits the requirements of a given task. A comprehensive overview for the use of XAI in climate science was recently given by the work of Bommer et al. (2024). The goal of this discipline is to shed light on the decision-making process. There are two basic approaches, called the model-aware and the model-agnostic methods. The former looks at the structures of the model itself (weights) while the latter primarily looks at the change of the output with respect to small changes in the input. This work only deals with the first approach, the model-aware methods, since the model-agnostic methods were not considered in this project. The so-called model-aware methods comprise all the gradient-based methods as well as layerwise relevance propagation (LRP), which follows a different approach. The theory behind the gradient-based methods is the computation of the local gradients, backwards from the output to the input. This allows us to estimate the ANN's sensitivity by changing the input a little bit and see how this affects the outcome based on the gradients (Baehrens et al. 2010). The positive or negative gradients all contribute to the final outcome with different quantities. There are various versions of this using this underlying principle. Since the raw gradient is typically quite noisy, many methods mainly aim for the reduction of the noise such that only the most salient features influencing the gradient are finally part of the result. These are methods such as SmoothGrad, NoiseGrad, FusionGrad (Bommer et al. 2024) and many more. Since the demands for explanations can vary as much as the demands for a CNN, many methods are constantly developed contributing to a dynamic, relatively new discipline.

It is important to note that these methods can be extremely useful for gaining insights into the decision-making process of a CNN classifier, but these explanations are obviously completely dependent on the accuracy of the CNN. Many issues such as overfitting have to be considered when analyzing for example heat maps obtained from a XAI method. Therefore it is important, since the explanations are lacking a ground truth as well, to evaluate them. Again there are various methods for doing that as well, as shown in the work by Hedström et al. (2023). Most methods aim for an iterative changing or perturbing of the input data based on the explanations. The CNN shall then predict to get an estimation of the influence of e.g. a certain pixel in a picture. By iteratively perturbing different regions of a given image while measuring the CNN's performance, it is possible to develop metrics which can deliver an estimation of the reliability for an explanation. It is especially important in the context of climate model evaluation to have a good understanding of potential explanations obtained from any given XAI method. As a first

step, it is really helpful having these tools at hand, but still there is a lot to consider before they can be properly used. Nevertheless, they appear to be promising methods for the use of machine learning in climate model evaluation and beyond.

The work by Brunner and Sippel (2023) provides an excellent entry point to the use of such XAI methods and forms the foundation upon which this thesis is built. As mentioned in 1.3, they were able to build a CNN capable of robustly distinguishing observational data from climate model data. Given the high accuracy of this CNN, important questions arise: how would such XAI methods apply to such a CNN? Are there geographical areas playing a more important role and what knowledge gain is expectable when applying XAI to such a CNN?
XAI therefore creates opportunities for gaining new insights into the various datasets in climamte science. Ideally, it could help in discovering unknown biases within the data such a CNN can learn from.

2 Research Questions

The overall goal of this work is to get an understanding of how these XAI methods apply in the context of climate science and how the results can be used for further examinations. This work is conducted in the Python programming language and makes use of various frameworks, most importantly those for machine learning and the XAI methods. The two central research questions are:

1. **Which geographical regions are most important for a skillful prediction with respect to daily temperature maps?**

   The results of layerwise relevance propagation provide a relevance score for every input feature the Convolutional Neural Network (CNN) was trained with. The input features in this case are the grid-cells of daily temperature maps. Therefore, this method should help in identifying the geographical regions that contribute the most to the CNNs decision when predicting "climate model data" or "observational data". Since the CNN classifier is very accurate, this can reveal where the observational based data and the climate model data differ the most or exhibit differences that can be systematically learned by machine learning methods like a CNN.

2. **To what extent is it possible to pinpoint the predictive skill of CNNs when identifying climate models?**

   In the context of this work, it shall be investigated how far it is possible to get insights into the reasons for the predictive skill of the CNN. Regarding all the potential sources of variation that a CNN can learn from in the context of climate data, it is assumed that the interpretation of XAI results is far from being straightforward.

The two questions are directly linked to the application of the XAI methods used here, addressing the issue of possible knowledge gain through the use of XAI techniques. It has been shown that many XAI methods work out pretty well, but what we can expect in the framework of climate science is not obvious at all. Furthermore, it is not trivial to interpret results generated by XAI methods. Given that this is a relatively modern subfield of AI, these questions have not yet been widely addressed in climate science.

3 Data and Methods

All the computations were carried out on the Jet server, a high performance cluster provided by the department of meteorology and geosciences at the University of Vienna. The jobs were executed by using jupyter notebooks, which also allow for a convenient sharing and explaining of the results. The python programming language was used (v. 3.9.0) throughout the entire work, making use of important frameworks. Most importantly Tensorflow (2.14.0) and iNNvestigate (2.1.2), which are the frameworks building up the core of the conducted experiments. For a list of all the versions of the frameworks used as well as the operating system see 7.

3.1 Climate model data and observational data

The data used in this work comprises daily temperature maps of all available CMIP-6 models (comprising 43 climate models) and four observational datasets.
It is the same setup used in the paper by Brunner and Sippel (2023). The whole preprocessed data was provided by Lukas Brunner.
The observational products used are ERA5, MERRA2, DOISST and 20CR. These cover different approaches and therefore some diversity within the observational products: ERA5, MERRA2 and 20CR are reanalysis datasets and DOISST is based on an interpolation method using in situ and remote sensing of sea surface temperature, not providing any data on land. It is a combination of ship, buoy and satellite Sea Surface Temperatures (SSTs) undergoing an interpolation procedure (Huang et al. 2020). For solving this, a land-mask is used whenever the DOISST dataset was part of the training or prediction processes. After applying the land-mask, the grid-cells (which are the input features in this scenario) are reduced to a number of 6888. Whenever the DOISST dataset was excluded from training, the number of input features resulted in 10368 grid cells due to the grid-cells on land which were used then.
Two different resolutions of the daily temperature maps were used as well. To get an intuition for results obtained by layerwise relevance propagation on a lower level of complexity and how they compare to a more complex setup, a coarser resolution of the data was additionally used in combination with a less complex classifier. Therefore, the maps were also regridded to 10.0°x 10.0°resulting in 2592 grid cells, having less computational costs for the methods used later during the process. Mainly, the 2.5°x 2.5°data was used to be able to first reproduce the results of Brunner and Sippel (2023) and to use the original classifier trained on the higher resolution for the explainable AI methods.
Two differently preprocessed datasets were used. First, temperature maps where the global mean was removed from all the temperature maps. This data is referred to as absolute historical (ABS) in this work. In addition, following the setup of Brunner and Sippel (2023), data with the mean seasonal cycle removed from all the temperature maps, referred to as deseasonalized historical data (DES). The mean seasonal cycle was computed by calculating the average over ±15 days around each day for all the years used during training. This was done for reducing the variance within the datasets, from which classifiers are said to mainly retrieve their information from. By subtracting the mean seasonal cycle, many dataset specific biases should be removed, making the prediction much harder for the CCN classifier. For the removal of the global mean, the global mean of each temperature map was computed and subsequently subtracted from the corresponding temperature map.

3.2 Convolutional Neural Networks

Two different CNN architectures were used in this work. The purpose was to compare different levels of complexity, overall results, and the impact of using various datasets.
First, a rather simplified model was set up to build up a pipeline for the XAI experiments. This CNN classifier for image classification has three convolutional layers with increasing filter sizes (64, 128, 256), each followed by a max-pooling layer. The model then flattens the feature maps and includes a dense output layer with a softmax activation for classification. It uses the Adam optimizer and sparse categorical cross entropy loss, targeting classification accuracy.
The classifier used by Brunner and Sippel (2023), which was more deeply examined later: it consists of 8 hidden layers as well as the input and an output layer. The overall architecture is the same as for the less complex classifier, which just has less hidden layers and one convolutional layer less. The two-dimensional daily temperature maps could easily be passed to the CNN, which was done by using a shape of samples $\times$ latitude $\times$ longitude $\times$ 1, where the last dimension represents the color channel for the maps, which is reflecting the temperature values in this case.
For the calibration of the classifier, not only the accuracy and the loss were considered, but also the confidence. The confidence was then used to compare against the actual accuracy, therefore comparing the final probability output score of the classifier for a given class (model or observation) with the true accuracy. This results in a so-called reliability plot. The method for plotting these reliability plots was already provided by Lukas Brunner. The results of these reliability plots are provided in the Appendix.
For the activations of the hidden layers, a standard activation function for CNNs, ReLU (rectified linear unit), was chosen. For the output layer, the softmax function was used, which normalizes the probability scores across the different classes. Detailed architectures of the two CNNs can be found in the Appendix.
For training and validating the more complex, deep classifier, a time period of 20 years was used for all the datasets, from 1982-2001. The CNN was tested on unseen data, namely from the time period 2005-2014. The less complex CNN classifier was trained on a specified set of models and only with 3 out of the four observations, excluding the DOISST dataset to obtain results on land. The time spans were the same though. The lists of datasets used for this classifier can also be found in the Appendix.
To complement the simplified model trained on the data with a coarser resolution, the deep CNN classifier was also trained without the DOISST dataset to have also higher resolution explanations on land. Therefore, three different CNN classifiers were trained in the scope of this work. These three classifiers were then used for investigating the important geographical regions which heavily influence the decision making process of the CNN classifier.
The two classes to predict are the observational based data or the climate model output. Therefore, so-called binary CNN classifiers are used that shall predict two different classes (climate model output or observational based data).

3.3 Layerwise Relevance-Propagation

Layerwise relevance-propagation (LRP) exploits the structure of neural networks by propagating backwards from the output scores to the input features. An underlying key-principle is the conservation rule (Montavon et al. 2019). This rule reflects the fact that the received activations of a neuron at a given layer must be redistributed to the lower layer. The activations, which are defined by the product between the weights of a neural network and the activation function (logit scores), are of different extent, reflecting the importance of certain features. Features to which the network assigns high relevance to, will be weighted more heavily. LRP propagates through the network collecting the logit sores at each node, the results of the weighted sums of inputs. The algorithm stops when the input layer is reached, therefore the final score for a certain input feature is obtained. This can be positive or negative, reflecting the influence on the decision towards a certain class.

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k \tag{1}$$

In this formula, j and k are considered to be the neurons of two consecutive layers (Montavon et al. 2019). The variable z reflects the relevance of neuron j on neuron k, therefore it is composed of the activations times the weights a neuron receives. This is the basic underlying formula for the LRP algorithm. However, in practice an additional rule or several rules are applied to the algorithm. Since the results for the ordinary formula are typically noisy, these rules aim for filtering out the most outstanding signals. Consequently, the rules typically apply in the denominator of the basic formula, such that weak signals diminish. An overview of rules can be found in the work by Montavon et al. (2019). Within this project two of them were tested (LRPZ, Epsilon rule) to find out that the results are nearly identical. In addition, the method of the Input Gradient was also tested to find out that it this led to the same insight. However, the Epsilon rule was chosen since it just adds a term to the denominator and doesn't exclude the negative signals, allowing results being a heatmap of important features influencing the decisions positively and negatively for the two classes in this binary classifier.

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k \tag{2}$$

The second formula showing the Epsilon rule in the context of the LRP algorithm also reflects for the composition of the relevance scores. These scores are the product of the incoming activations a neuron receives from the outputs of the previous layer and the assigned weight, which is constantly adapted during the learning process. Through these layerwise application of the rule, it is possible to trace back the input features that are most relevant for the decision of any classifier. It is also possible to apply different rules to different layers, known as the composite LRP. This was also tested, but didn't show any difference regarding the results as well. For this reason, the Epsilon rule was mainly used throughout this work.

The framework chosen for making use of the LRP algorithm was iNNvestigate (Alber et al. 2019). In previous literature it has been reported to work well for tensorflow models (Bommer et al. 2024), employed here as well. Moreover, the choice is anyway restricted by the machine learning framework used (Pytorch, Tensorflow etc.). There are a couple of frameworks out there such as Captum, Zennit, tf.explain to name the most prominent ones, but many are weakly documented or only work for a specific framework such as e.g. Pytorch. For these reasons, iNNvestigate was chosen for the well documented methods and the previously successful application in the context of climate science.

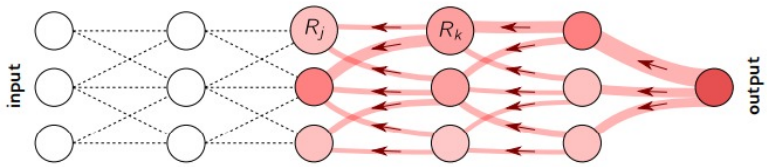


Figure 2: *Schematic of the LRP process. Propagates from the output layer to the input layer. At each neuron, the received relevance scores are redistributed to the neurons of the lower layer. Source: Montavon et al. (2019), license number: 5858670090928).*

3.4 LRP Evaluation by Occlusion

To evaluate the explanations generated by LRP, a method was used where highly relevant areas are occluded. The assumption is that occluding areas identified as highly relevant by the LRP algorithm should cause a more significant drop in the classifier's performance compared to using a random mask that covers randomly chosen grid cells. This approach also seemed to be interesting in the future for easily evaluating geographical regions that play a crucial role in the identification of climate models or observational data. This method is often referred to as occlusion sensitivity in the literature. The simple approach chosen in this work is to define a fraction that shall be occluded. This fraction shall be increased iteratively to compute the accuracy afterwards. The expectation would be that the accuracy drops considerably more for the relevance based case than in the scenario of using random occlusion, assuming that the explanations are valid. A method for randomly occluding grid cells in the daily temperature maps was implemented as well to be able to compare it to the relevance based occlusion (see 4.2.6).

```
def create_relevance_mask(data, fraction_percent):

    # Flatten the input array
    flat_data = data.flatten()

    # Calculate the total number of elements to be masked
    n_elements = int(len(flat_data) * (fraction_percent / 100) / 2)

    # Extract indices of the top n_elements highest values
    top_indices = np.argpartition(flat_data, -n_elements)[-n_elements:]

    # Extract indices of the bottom n_elements smallest values
    bottom_indices = np.argpartition(flat_data,
        n_elements)[:n_elements]

    # Initialize a flat mask with False values
    flat_mask = np.zeros_like(flat_data, dtype=bool)
```

```
17
18      # Mark the top and bottom indices as True in the flat mask
19      flat_mask[top_indices] = True
20      flat_mask[bottom_indices] = True
21
22      # Reshape the flat mask to the original data shape
23      mask = flat_mask.reshape(data.shape)
24
25      return mask
```

Method for creating the masks for the relevance based occlusion. Grid cells are occluded which were found to be most relevant by the application of the LRP algorithm.

This function uses NumPy's argpartition function, which efficiently partitions the array without fully sorting it, placing smaller numbers in front of a specific pivot element. The fraction as well as the data can be passed to the function, which then divides the fraction by two to have half of the fractions for the highest relevance scores in magnitude and the other half for the lowest. The random occlusion was implemented almost the same way except for using the random choice function by numpy to determine the random grid cells of the array. Also, the fraction was not split, since this is not necessary in the random case.

3.5 Pearson Correlation

The Pearson correlation is used (see 4.2.4) to compute the correlations between the relevance scores that make up the explanations. The goal was to compare them between the case where only the global mean was removed and where the mean season cycle was subtracted in addition. The underlying assumption for using this method would be, that the explanations in the case where the mean seasonal cycle was removed should exhibit spatially more complex patterns, leading to more independence between the single input features (grid cells). The Pearson Correlation is a normalized covariance measure, defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{3}$$

So the Pearson Correlation is basically the covariance of any two variables divided by their corresponding standard deviations (J. 2016). The result ranges from -1 to 1, depending on the relationship. Negative correlations would mean that they are inversely proportional while positive ones would mean that they are directly proportional. A coefficient of zero equals no correlation at all. Here, the implementation by Scipy is used (pearsonr).

4 Results

The final outcome of LRP is a relevance map having the same shape as the input. These are the so-called explanations. This is a big advantage of the method, exploiting the existing structure of the CNN without modifying anything of the architecture. As previously described, those resulting relevance maps reflect the weights the classifier has assigned to each grid cell in the temperature maps during the training process. More precisely, the outcomes are the logit scores, therefore the activation function multiplied with the result of each regression task at each neuron. In the following section, the outcomes of LRP as well as further analysis is provided.
To achieve clarity for the reader, some terms shall be explained. As mentioned in the methods 3.1, the data was preprocessed in two ways. The data where only the annual global mean was removed is referred to as the absolute historical data throughout this section. The data where the the mean seasonal cycle was removed in addition is referred to as the deseasonalized historical data. The term classes or class refers to the different datasets the trained CNN classifier predicts (climate model output or observational based data). The results obtained by LRP deliver the explanations of the decisions for the two different classes the CNN predicted.

Since some terms are used at a high frequency within the following section, a few acronyms will be introduced for a better readability:

- **ABS** – Absolute Historical
- **DES** – Deseasonalized Historical
- **MOD** – Climate Model Output
- **OBS** – Observational Based Data

This section starts with a simplified example that was part of the first experiments. This setup shall introduce the main aspects of the obtained explanations, their interpretation as well as implications. Important dynamics are already observable for this simplified example, some important geographical regions are already identifiable as well. After that, results for the different deeper CNN classifiers are provided as well as results of further statistical analysis.

For describing important geographical regions, the terminology of the AR-6 regions are used (Iturbide et al. 2020).

4.1 A Simplified Setup

The first results of this work are those for a CNN classifier comprising less hidden layers than the deep classifier created by Brunner and Sippel (2023), which was used later as well. Moreover, only a part of the climate models participating in CMIP-6 and 3 observational products are used (see Appendix 7). This experiment should give a first insight into whether a less complex classifier is comparable to a more complex one, regarding the classifier itself and the resolution of the data. Finally, the explanations obtained by LRP shall give insights how this lower level of complexity affects the CNN's decision. This can also be seen as a testing step to see what can be captured by using a lower level of

resolution. Using this set up, it was additionally possible to get a first intuition how this method works while keeping the computational demand low. The data used here were 2-dimensional temperature maps having a resolution of 10.0˚x 10.0˚grid cells, resulting in 2592 grid cells, where each one represents an input feature. For the ABS data, the skill was quite remarkable with an accuracy of over 99% (7). For the DES dataset, the skill decreased significantly with an accuracy of 86%. In addition, this classifier was pretty overconfident and the accuracy could only be reached by using twice the amount of OBS during the training process, which could have caused overfitting. For this toy model set up, the small goal was to see how differences within the two preprocessing setups emerge at this resolution. This observation of decreased skill is, to some extent, reflected in the explanations for the classifier (see Figure 3). While for the ABS data the explanations are showing a strong inverse pattern, meaning that regions identified by layerwise relevance propagation (LRP) as positively influencing the decision towards one class, negatively affect the decision towards the other class. Regions or grid cells exhibiting a highly positive or negative score are highly relevant for the decision making process. For the DES case though, this clear pattern can't be observed anymore. The negative correlation between the two explanations of the OBS and MOD still hold in the DES case, but it is not strongly exhibited. The classifier does not distinguish as accurately as in the ABS case. In the explanations, one can observe that the classifier includes more grid cells into the decision process: while in the ABS case the number of pixels not having any considerable importance assigned is relatively high, leading to very concentrated explanations, in the DES case this number is seemingly low. The CNN classifier distributed the relevance scores more evenly across the map, more broadly covering different geographical regions than being strongly focused on just a few.

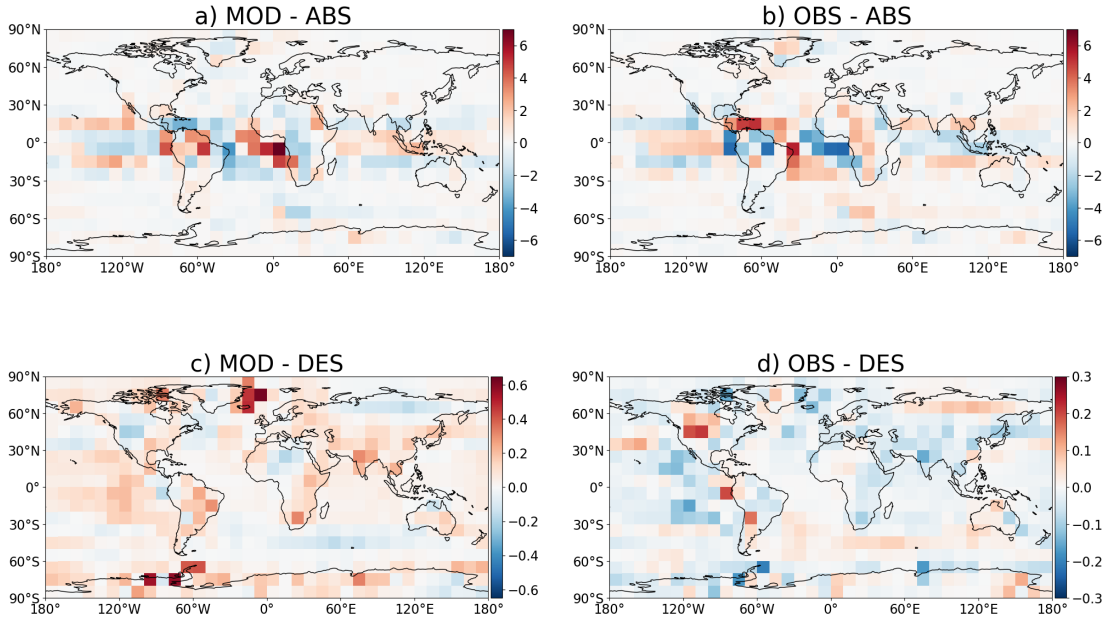


Figure 3: *Relevance maps showing the weights per grid-cell generated by LRP-epsilon rule. 2400 samples of MOD and OBS were used in combination with a trained simplified classifier having three convolutional (6 hidden in total) layers.*

Looking at the relevance maps in the DES case, it is possible to observe a more complex spatial pattern as in the ABS case. There are no clear hot-spots being weighted

much higher. Overall, the relevance scores are indeed inversely related, but not showing geographical regions that are clearly highly relevant. Since the classifier used for these explanations was very overconfident, it is the question, whether these results for the overconfidence are due to overfitting. Overfitting is always an issue when using CNN's and has to be considered, although it is not always straightforward to prove. Nevertheless, the effect of subtracting the mean seasonal cycle from the data is clearly reflected in the explanations, leading to a distribution of relevance scores, which are not as strongly concentrated as in the ABS case. Regarding the reduced skill in the DES case, it makes sense that the explanations lack geographical regions that are highly relevant since the classifier was obviously not able to derive clear differences between the classes to the same extent as in the ABS case.

4.2 Deep CNN Classifier

The classifier used in the scope of this work comprising 9 layers (8 hidden layers + 1 input layer) is referred to as the deep classifier, which is also used by Brunner and Sippel (2023). It is the classifier capable of robustly distinguishing MOD from OBS for all dataset types. The question being tackled would be, what geographical regions as well as spatial patterns might be extracted by the classifier for conducting the decision with such a high accuracy. Since the accuracy is very high for both types of preprocessed data (DES and ABS), the goal is to find out what geographical regions and spatial patterns influence the decision most heavily in both cases. Additionally, the effect of deseasonalization on the explanations is worth investigating.

The following section will start with explanations generated by LRP for the different CNN classifiers, comprising the explanation maps of the CNN classifier trained without the DOISST dataset as well as the one where it was included. The results will be compared and further examinations are conducted to get an insight into the behaviour of the CNN. This also includes the investigations of the misclassified samples as well. Finally, the explanations are evaluated by making use of an occlusion procedure where the grid cells, which were defined to be most relevant, are iteratively occluded before prediction. The section will be completed with explanations of a classifier that was also trained on occluded data.

4.2.1 Trained without DOISST

To provide explanations on land with a higher resolution and complement the explanations of the toy model, the following explanations are provided for the full classifier trained on three reanalysis (MERRA2, ER5, 20CR) datasets and all the available CMIP-6 models. It can be seen as the next step to investigate differences between the two resolutions and two classifiers with different performance: the full classifier, trained on the mentioned data, is highly accurate for both types of preprocessed data, also having a tendency to be a bit overconfident in the DES case. The overall accuracy exceeds 96% though, showing that the vast majority of samples was correctly classified in the DES case. Since the higher resolution leads to a much clearer pattern, specifying the important geographical regions (see Figure 4), it can be observed that the explanations of the lower resolution indeed share the same tendencies for some highly relevant regions. Nevertheless, the higher resolution obviously leads to spatially more accurate results, revealing

relevant geographical regions that can be more clearly identified.

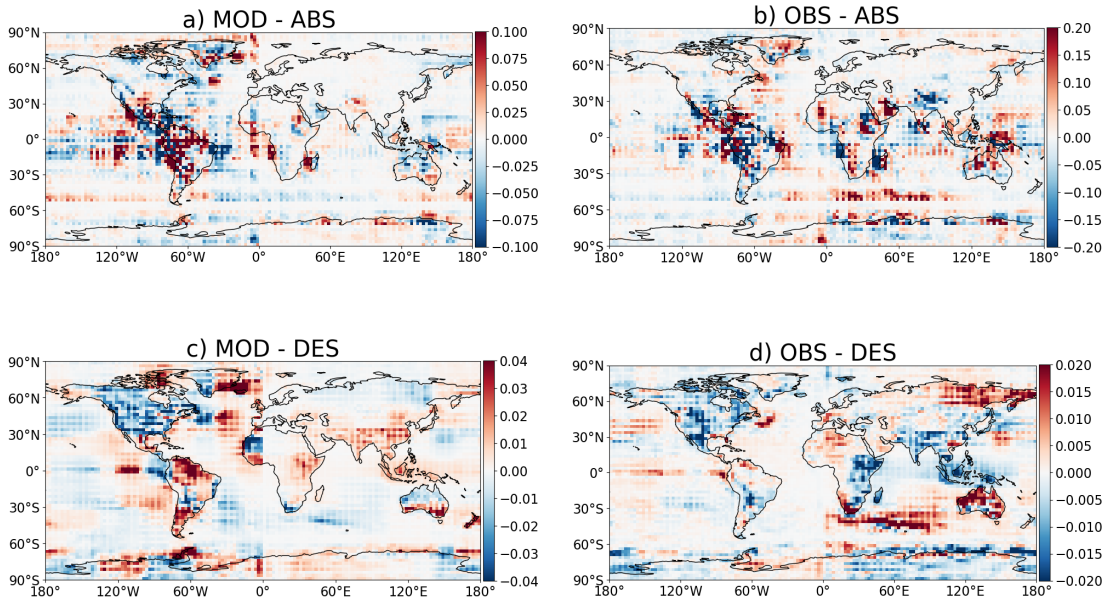


Figure 4: *Relevance maps showing the weights per grid-cell generated by LRP-epsilon rule. For generating the explanations, 1000 samples per dataset were used, resulting in a total amount of 46000 samples, 43000 for the MOD and 3000 for the OBS. Only reanalysis data used for the OBS. CNN with 8 hidden layers.*

Again, as for the toy example, the explanations show a very strong inverse pattern in the ABS case. Although the highly relevant grid cells are strongly accumulated around South America,their positive and negative contributions are very mixed within a pretty confined space. Obviously the bias contained in the ABS case provides enough information for the classifier to make the distinction very accurately. A lot of regions are not even considerably weighted, while a small part of the temperature map is heavily influencing the CNNs' decision. Interestingly, positive and negative grid cells in the explanations of the ABS data are spatially mixed in and around South America, raising the question of how the bias in the data can be separated in such a confined space.

The effect of deseasonalization has a strong impact on the explanations. As already observed in the toy model (Fig.3), the strong inversely related explanations can't be observed anymore. At this resolution, most regions are indeed inversely related with respect to the relevance scores, but there are very prominent regions on the map sharing the same type of contribution (positive or negative) among the explanations for both classes, such as North America and parts in South America. Moreover, the South Atlantic region seems to play a decisive role for the decision towards the OBS, while this region is not influencing the decision that heavily in the case of the MOD. The fact that North America shares the same tendencies in the relevance scores for both classes (MOD and OBS), raises the question how this large part of the map is not influencing the classifiers performance that badly, regarding the overall high accuracy. Looking closer at the explanations, it becomes clear that the spatial arrangement of the relevance scores is different. The question here would be, how the CNN is capable of capturing the differences in the spatial pattern of the relevance scores, independently of the actual score (positive or negative). Generally speaking, another result of the deseasonalization is that simply more grid cells become more relevant for the decision. Especially in the explanations of the OBS, the relevance map is pretty filled, also having highly concentrated areas which seem to be very important for the decision, such as the Southern Ocean for the OBS. The DES explanations of the MOD indeed show clear regions of higher importance such as the North Atlantic or Greenland/Iceland or northern South America. But again, don't exhibit such highly concentrated areas of relevance scores, since the map is more evenly covered with relevance scores compared to the ABS case.

4.2.2 DOISST Dataset included

The following explanations provided are all obtained by using classifiers where the DOISST dataset was included in the training process. The classifier used is exactly the same one used by Brunner and Sippel (2023). After having gained important insights into the mechanism of layerwise relevance propagation through a simplified example, also testing the influence of different resolutions and CNN complexity on the CNN's performance, the following set up shall deliver more precisely the geographical regions most important for the distinction between MOD and OBS. Since this CNN classifier has the highest skill compared to the others, it can be expected that the explanations should be of higher quality with respect to the localization of important regions. Nevertheless, the results will largely confirm the rough geographical tendencies of the lower resolution data, but the explanations in the DES case definitely increased in quality. For the following results, it is indeed possible to locate several highly important regions for both the differently preproceseed data. It is also interesting to compare the results to the explanations on land, how they diverge or which patterns persist. Additionally, the influence of another dataset being fundamentally different compared to the other OBS is worth investigating. It shall be emphasized that for the explanations in 4.2.1 only reanalysis datasets are used to represent the class of OBS.

ABS Data

In this scenario, another dataset was included into the training process, the Daily Optimal Sea Surface Temperature (DOISST), which can be seen as the more direct OBS compared to the reanalysis datasets.
The CNNs performance increased with respect to the accuracy (see 7). This would make sense, since one more dataset is included, expanding the class of OBS and the learnable feature space. Furthermore, the dataset is very different from the other three reanalysis datasets, which is also observable in the explanations (see 4.2.3). Therefore, the CNN can rely on more distinct features for the class of OBS, when making a prediction. Regarding the reliability, both deep classifiers showed very similar behaviour. Here, since the DOISST dataset does not provide data on land, a land mask is used to include all the OBS into the training process. Zeros were introduced into the grid cells on land. This is of course a straightforward approach, but raises the question how neighboring grid cells share dependencies or similarities and how this might affect the learning process of a CNN classifier as well as the corresponding explanations.

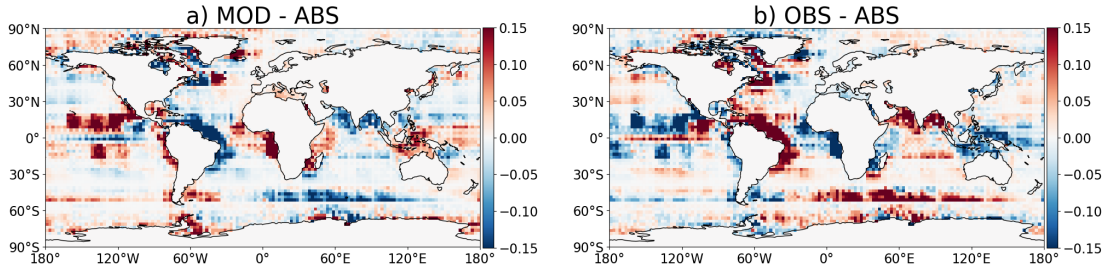


Figure 5: *Explanations for the deep CNN classifier trained with DOISST data. Land data excluded. 1000 samples per dataset taken for the LRP algorithm resulting in 47'000 samples in total.*

For the masked explanations where the DOISST dataset is included 4.2.2, coastal regions seem to increase in importance. Especially a few coastal regions such as northern South America, western South Africa up to western Africa, including parts of Central Africa as being the areas exhibiting the largest scores in magnitude. The Western African region could already be observed for the lower resolution (4.1) and seems to be a region where the CNN classifier is able to extract distinctive biases among the two classes of datasets already at lower resolutions. Moreover, in the explanations where land was included, these regions were also found to be highly relevant. Nevertheless, at the coast of Western Africa, the scores are more concentrated compared to the explanations without including DOISST. The Southern Ocean area appears to be highly relevant for both classes, while this was not the case when DOISST was excluded. The Southern Ocean considerably influenced the decision in favor of OBS, particularly in scenarios excluding DOISST (ABS and DES data). When DOISST is included and land data is excluded, relevance scores notably increase in coastal regions areas where land scores were initially high. Additionally, the eastern Pacific assumes a more prominent role. The Southern Ocean positively impacts decisions based on OBS, whereas the Pacific Ocean tends to sway decisions towards MOD. But the equatorial Pacific shows highly positive relevance scores for OBS, indicating its higher relevance for decisions favoring OBS in the context of the ABS data.

This pattern also aligns for example with the coefficients learned by a logistic regression classifier created by Brunner and Sippel (2023), for a direct comparison see Appendix (7). Interesting is the change regarding the signs of the relevance scores from positive to negative and vice versa. For instance, the Arabian Sea, Bay of Bengal and South Asia are positively scoring for the MOD class in the explanations on land, while they are then negatively promoting the decision towards the MOD when using the land mask. Furthermore, another interesting spatial occurrence are the well ordered relevance scores for the explanations where DOISST was included: while South America was very messy regarding the distribution of positive and negative relevance scores, meaning they appeared spatially mixed and close together, this is completely different for the explanations in the land masked case. In South America, where the northern coast plays a decisive role for the decision, the scores share the same sign. While e.g. the positive scores distribute over the east coast of South America, the west coast holds the negative scores considering the class of OBS, for the MOD this is the complete opposite. The effect of excluding data on land as well as additionally using one more dataset providing more data on sea leads to more globally distributed relevance scores, while the relevance scores for the explanations without DOISST are generally strongly located within the lower latitudes (4.2.1), most of them ranging from 30° South to 30° North. Except for the Southern Ocean region, which seems to be especially relevant for the decision towards the OBS.

DES Data

The effect of deseasonalization is again strongly visible. While some areas almost vanish compared to the ABS case, other regions become increasingly important or don't exhibit such a large difference in the relevance scores. For example, it seems as if the seasonal cycle within the pacific delivers a lot of bias the classifier can learn from. While this area (N. Pacific Ocean, E. Pacific Ocean and S. Pacific Ocean) concentrated many high relevance scores in magnitude for the ABS data, this doesn't apply to the DES case. In the explanations for the DES case the E. Pacific Ocean remains to be of higher importance for the decision promoting it towards the MOD. The coastal area for S. America is not that strongly affected by the removal of the mean seasonal cycle, but interestingly the North Atlantic (NA) takes a more dominant role.

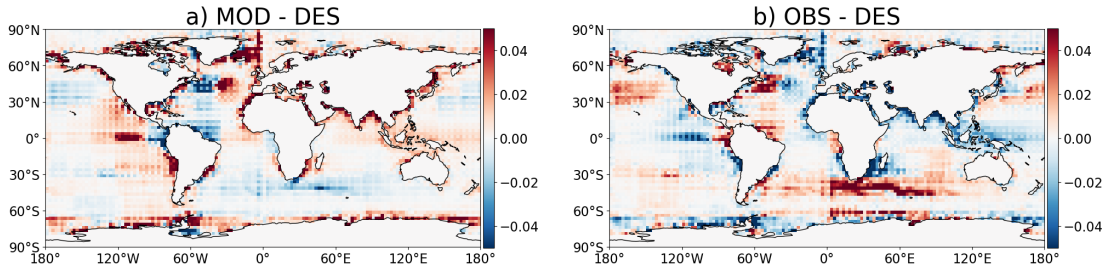


Figure 6: *Explanations for the deep classifier in the DES case. The removal of the mean seasonal cycle leads to a different spatial distribution of relevant grid cells.*

Regarding all the explanations so far, it is true that the north Atlantic region shows up everywhere, but always clearer visible in the DES case. However, since this area persists throughout the explanations in varying magnitude, this could be an interesting example worth investigating at a deeper level. Another area which is constantly influencing the CNN's decision to a large extent is the southern Ocean, again especially for the class of OBS. It is true that it also plays an important role in the ABS case, where the area is even more balanced between the classes, regarding the magnitudes of the relevance scores. But it is another interesting effect of the deseasonalization, leading to the CNN assigning very high relevance to this area when deciding for the OBS. The spatial distribution of relevance scores is quite different between ABS and DES data, but still in both cases they are confined within the southern Ocean. Nevertheless, it raises the interesting question, why the southern Ocean increases so heavily for the decision in favor of the OBS, while the relative relevance is reduced in the class of MOD. Another observation that can be made, is the very drastic change of relevance scores within Arabian Sea, Bay of Bengal and South Asia. The region flips the signs of the relevance scores upon deseasonalization. While the relevance scores are always the opposite of the south-eastern Asia region in the ABS case (4.2.2), all these regions share the same positive/negative relevance scores (depending on explanation of OBS or MOD) in the DES explanations. To mention the extremes at last, the Greenland/Iceland region also exhibits strongly concentrated areas of high relevance for the distinction, also a bit more relevant in the DES case. Still, this area seems to be of higher relevance also when using the ABS data to train the CNN classifier. The Antarctic region often locates many grid cells of higher relevance along the coastal regions exhibiting a band-like structure across the globe reaching across western Antarctica (WAN) to eastern Antarctica (EAN). This pattern is well established in the DES case.

4.2.3 Differences Among Observational Datasets

Analyzing these averaged explanations with respect to each predictable class or using them as a first step for understanding the decision making process of a well performing classifier indeed gives a good intuition about the overall tendency. But also for the case of a binary classifier, the explanations might differ substantially for different datasets from the same class. Regarding the binary classifier, it can be stated that the explanations are relatively homogenous (see 7) regarding the different climate model outputs.

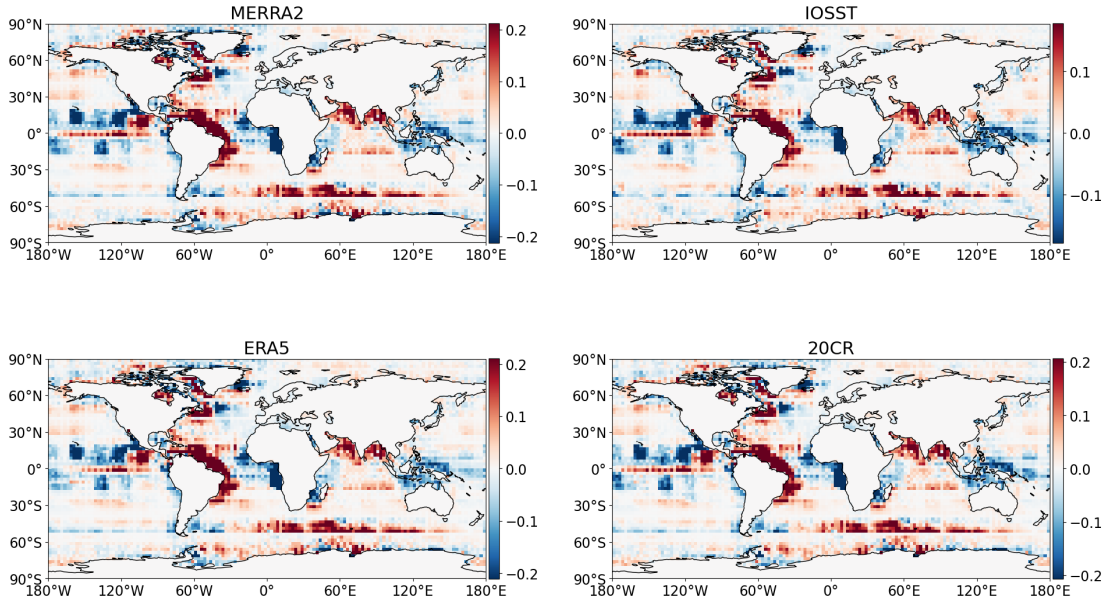


Figure 7: *Explanations per OBS where the average was taken per dataset, ABS (annual global mean removed). 1000 samples per dataset used for computing LRP. Only minor variance among the different datasets observable.*

For the OBS though, this can be argued for the 3 reanalysis datasets, while the explanations of the DOISST dataset are quite different in the DES case. To visualize this, the plots for the class of the OBS are provided in the case of the ABS and the DES case. It is interesting that the CNN doesn't adapt its weights so much to the different datasets among the same class in the ABS case. Again, the question comes up how the underlying biases for the CNN trained on ABS data are sufficient to make an accurate prediction without the need to learn more dataset specific patterns. Since the explanations only slightly differ in magnitude for a few grid cells, there is obviously no need for the CNN to exploit additional patterns (e.g. spatial patterns) from the data which could lead to the extraction of additional information according to the different datasets. Since the weights are optimized with respect to the loss function of a CNN, meaning that it adjusts its weights until the error is minimized, it seems as if the CNN doesn't have to rely on complex spatial patterns in the ABS data to achieve an almost perfect accuracy. It is obviously the case that a few geographical regions expose enough distinct patterns that are easily accessible for the CNN. These areas seem to contain enough differences between the biases of the OBS and the MOD, suggesting that the climatological bias seems to be more obvious for instance. For the DES case though, the difference between the DOISST dataset and the other three reanalysis datasets is very clear. Interestingly, a systematic

cold bias for the DOISST v2.0 in the indian ocean, which was used for training this CNN, was reported by Huang et al. 2021 Huang et al. 2020.

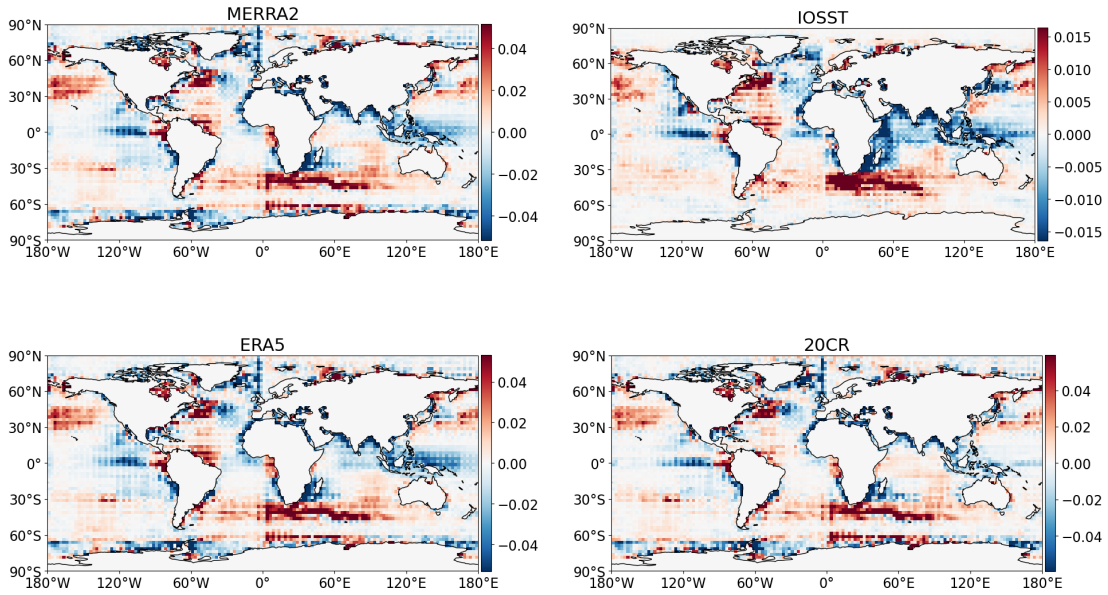


Figure 8: *Explanations per OBS dataset where the average was taken per dataset, DES (mean seasonal cycle removed). 1000 explanations were used per dataset for computing the LRP results. Higher variance among the different datasets.*

However, an obvious difference between the three reanalysis datasets and the DOISST dataset is reflected in the explanations. Overall, the indian ocean and the surrounding coastal regions receive more relevance, influencing the decision negatively towards the OBS. The coastal region of Eastern Africa exhibits much more relevance compared to the reanalysis datasets as well, negatively influencing the decision for the OBS. The Southern Ocean seems to play an important role as well, but the spatial distribution of the relevant grid cells in that region differs as well for the DOISST dataset.

Such XAI methods can indeed be very useful to get a first hint towards potential biases, but still there are too many in the data to make any conclusion. Observational uncertainty due to sampling bias, instrumental bias or historical data bias through alterations in the methods of data collection are far from trivial to determine. But these are all patterns a deep CNN could learn to some extent. It would always require additional investigations of given areas as the indian ocean which exhibits much more relevance in the explanation map of the DOISST dataset for example. It is interesting though, that these differences already emerge for a binary CNN classifier, which shows that also in the binary case, only distinguishing between MOD and OBS, the CNN is learning distinct patterns among the same class. Of course there are differences in the explanations of climate model outputs as well (see 7). To examine those differences more precisely though, it would be necessary to apply XAI methods on a multiclass CNN, which actually learned to distinguish between the single MOD datasets (or OBS). In such a scenario the CNN is forced to learn dataset specific patterns, which could then be disclosed by making use of LRP. This could lead to an additional knowledge gain with respect to geographical regions that may contain unknown systematic biases among the MOD or OBS themselves.

4.2.4 Statistical Insights into Explanations

Knowing the mean state of the explanations gives important tendencies, but it is far from trivial to investigate the identified areas further for an actual knowledge gain. Despite the fact that such a CNN acts as a black box not being transparent during its learning process, it is not even straight-forward to determine a rank within the identified regions by LRP. Of course, the obtained maps reflect the importance of certain grid cells for the CNN's decision, but again we are missing any reasoning. However, it is still not possible to really gain insights into the decision-making process, which keeps being enclosed. Evaluation techniques rely on the perturbation of the input (see 4.2.6) and always use the same classifier. So we are always applying metrics which compare the CNN's performance to itself when perturbing the input. This makes total sense, but still is only a relative measure, not showing why e.g. a certain region is more relevant. To tackle the fact that the results generated by such XAI methods may only give a first intuition of the underlying reasons, it might be reasonable to apply more classical statistics to a collection of grid cells based on the relevance scores for instance. As already mentioned, the effect of deseasonalization always exhibits the same pattern, namely leading to explanation maps where the relevance scores for the grid cells are more evenly spread across the map, leading to more grid cells contributing to the decision. Moreover, the explanations are not that strongly inversely related compared to the ABS data. This led to more complex spatial patterns in the explanations, not strongly focusing on relatively few areas anymore which seem to be mainly considered for the prediction by the CNN. Such observations can indeed be visualized, but the suspicion of more complex patterns which are captured is pretty hard to prove.

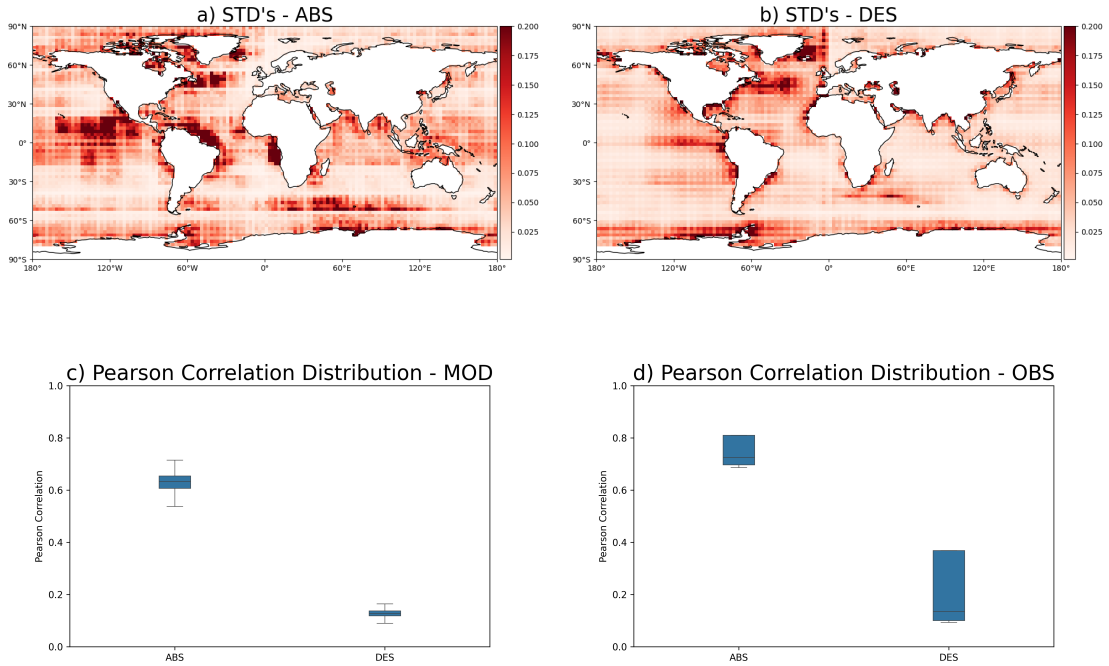


Figure 9: *a),b) show the map of standard deviations for the ABS explanations and the DES explanations. The boxplots c) and d) below show the distribution of the correlation coefficients for the MOD and the OBS.*

When examining the correlation of the explanations (Fig.9) to assess the linearity of the

relationship between individual relevance scores, it is evident that in the ABS case, the explanations exhibit a fairly strong correlation. This can be an indicator that the CNN relies on similar features across different samples. Regarding the previously observed differences among the classes (OBS and MOD) in the ABS case, where the maps were almost the perfect inverse every time, this seems reasonable. Another observation supporting the assumption that the same patterns are uniformly shared among the same classes is that the explanations were very homogenous in the ABS case.
On the other hand, the individual relevance scores, in the DES case, only share weak correlations. This strengthens the suspicion that the individual grid cells alone are not relying on the same or very similar information or biases. Rather, the CNN has to take different spatial arrangements into account, since for the two predicted classes, some grid cells overlapped with respect to the type of contribution (being negative or positive for both classes).

There are many more potential approaches to take a further look into the distributions of the explanations, the temperature maps themselves or only regions of the temperature maps. But this would go far beyond the scope of such a thesis. Nevertheless, the question of important grid cells is worth presenting since the actual determination is not completely obvious only by having such LRP maps at hand. Looking for example at the standard deviations of the explanations (see Fig.9) it is possible to narrow down the amount of grid cells which are highly relevant for the overall decision, containing the relevant areas for both the climate model output and the OBS. Since the relevance scores are ideally inversely related, areas of very high importance should exhibit a larger standard deviation. This finally captures geographical areas where relevance scores were the highest in magnitude, exhibiting high positive as well as negative absolute values for each class. However, the map best captures regions where scores were strongly inversely related. For example, the Southern Ocean is highlighted as highly important for the decision towards the OBS, particularly in the DES case. This region does not show a high standard deviation because the explanations did not have significantly high positive or negative relevance scores for the model's explanations (see Fig.6). So the standard deviations clearly captures the important regions in the case of the ABS data, but for the DES data, the result shows less important regions. However, this again shows that the relevance scores for both classes are not strongly inversely related for the DES data as observed and that the classifier exploits different spatial patterns in the data for both classes, the ABS and the DES data.

Another aspect that might be interesting to obtain additional insights into the classifiers' decisions is to look at the wrongly predicted samples. These are quite rare, but maybe disclose some dynamics of the classifiers decision making. For instance, upon deseasonalization, how do the misclassifications increase, are e.g. the OBS relatively overrepresented ? The following shall present the findings collected during the study of the misclassified samples.

4.2.5 Explaining Misclassifications

The misclassifications of the deep CNN classifier mainly happen in the DES case, where the number of samples wrongly predicted per dataset strongly increases for the OBS,

compared to the ABS case. The accuracy for the deep CNN classifier in the ABS case exceeds 99%, therefore only a few predictions are wrong. Nevertheless, the OBS are never confused with climate model data using the ABS data. Only climate model output mixed up with OBS in the ABS case. Since the accuracy for the deep classifier decreased a little bit in the DES case, there is almost a wrong prediction for every dataset for a reasonable amount of test samples. Looking at the overall accuracies alone might not reveal any big difference, since the deep classifier in the DES case shows an accuracy of over 98%. However, the absolute number of wrong predictions is eighty times that of the ABS case in the conducted experiment. For the predictions, 1000 samples per dataset were used, corresponding to 47'000 samples in total. The misclassified samples were extracted for the ABS and DES case for the deep classifiers used: one which also provided explanations on land and the second one only on the sea because of the use of the DOISST dataset. In addition, the effect of occlusion before the training process on the misclassifications was also examined (see 4.2.7).

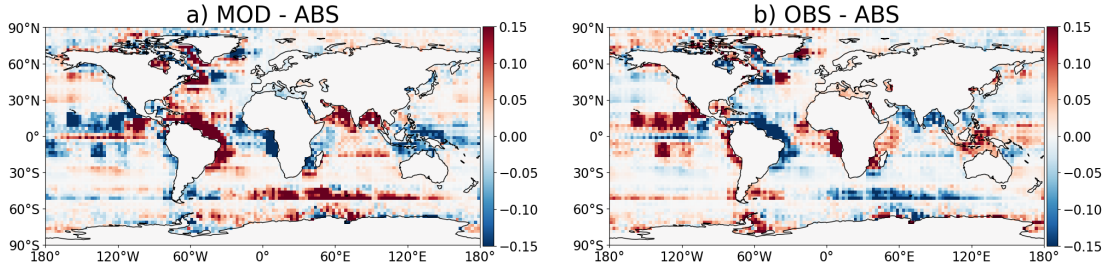


Figure 10: *Opposite relevance maps for the misclassified samples. Misclassified samples show the same relevance maps, just for the wrong class. The wrong predictions were extracted and the LRP algroithm applied to these samples.*

First, after the extraction of the misclassified samples, they were used to compute the relevance scores using LRP. So the same procedure for computing the relevance maps was applied to the wrongly predicted samples. It turns out that the relevance maps just flip the sign of their relevance scores, just swapping the relevance maps for the corresponding class such that it is the complete opposite for each class compared to the right relevance maps (see Fig. 8).
The explanations of the misclassified samples do not reveal much, except for the fact that the CNN classifier assigns the same relevance scores in magnitude. Or to put it differently, if the binary CNN classifier is wrong, the resulting explanation containing the relevance scores is the exact opposite compared to the explanations of the correct predictions. In the case of misclassified samples, the CNNs decision appears to reach a tipping point, resulting in the relevance scores flipping entirely. This produces a relevance map that is the exact inverse of the one associated with correctly classified samples, corresponding to the opposite class. This does not provide much information, since the relevance maps are almost exactly the same as the already obtained ones.

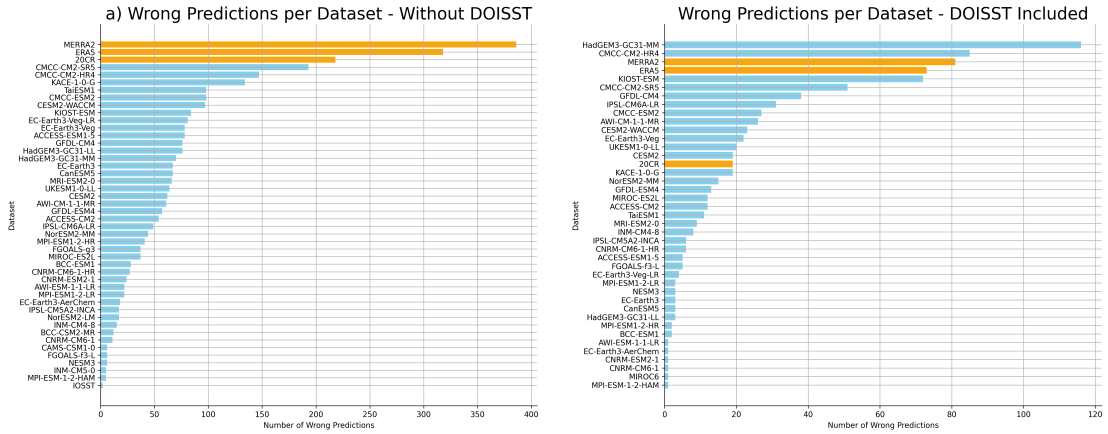


Figure 11: *Misclassifications for the deep CNN classifer trained on the DES data. Figure a) shows the wrong predictions for the classifer trained without the DOISST dataset while in b) this dataset was included. In the ABS case, both classifiers rarely misclassified any samples.*

However, when looking at the numbers each dataset was wrongly predicted, it can be observed that the effect of deseasonalization leads to a stronger increase in the misclassifications of the OBS. Every dataset of the OBS appears in the wrong predictions except for the DOISST dataset. This interesting finding aligns with the fact that the explanations obtained for the DOISST data are the most unique ones among the OBS. For the deep classifier, the DOISST samples had never been wrongly predicted for both the ABS and DES data. Another interesting observation is the increase of misclassifications for all the OBS in the case of the CNN classifier where the DOISST data was not used for training. The overall accuracy is generally 2% lower in the DES case, which indicates a decrease of skill, but it is probably worth looking at the specific datasets which increased the most. Doing that, one finds for example that MERRA2 has five times more misclassifications compared to the deep CNN where DOISST was included. Indeed, there are models as CMCC-CM2-HR4 or HadGEM3-GC31-MM which show pretty high numbers of wrong predictions in both cases, but not such a multiple of the misclassifications. Also, the wrong predictions for ERA5 are much higher for the classifier trained without DOISST data, therefore using reanalysis data only. These numbers were obtained by letting the classifier predict multiple times and use the rounded average. The numbers varied very little during the experiments, the relations stayed the same.

While these numbers should be taken with caution due to the difficulty in accurately estimating the variance of the classifiers' predictions, the accuracy combined with the misclassification rate suggests that by incorporating a more distinct OBS dataset into the training process generally makes the predictions easier for the CNN classifier. However, during the experiments, no big differences in the predictions could be observed regarding the misclassifications. While the explanations revealed a clear difference between the DOISST data and the data of the reanalysis datasets, the misclassifications show that only using reanalysis datasets, the predictions get more difficult for the CNN. Of course, the overall accuracy stays pretty high, but comparing e.g. the wrong predictions for MERRA2, the accuracy actually drops significantly: for the classifier that included

DOISST, the number of wrong predictions remained at around 80 for all runs out of 1,000 samples, whereas for the classifier trained solely on reanalysis datasets and climate model outputs, the number of misclassifications was around 400. MERRA2 misclassifications were observed to be the most severe ones for the classifier not using DOISST for training, but the other reanalysis datasets also increased by a multiple compared to the full set up including DOISST (Fig.11). On the other hand, it is interesting how the model HadGEM3-GC31-MM was more often misclassified by the CNN which has seen DOISST during training. In total, the amount of misclassifications was still almost half of the amount for the classifier trained with DOISST data.

4.2.6 Evaluation by Occlusion

To evaluate the explanations one has the opportunity to choose from a variety of tools, which are mainly focusing on the perturbation of the input. In this work, a straightforward method was implemented to test the reliability of the explanations. The implementation is also based on a perturbation of the input data.
Grid cells which were found to be highly relevant, are occluded before letting the CNN predict. The implementation provides different approaches for the occlusion values, which are typical in this domain: insertion of zeros, mean values or random noise. In the scope of the experiments, only the insertion of zeros was used, since the land mask used for the deep CNN also made use of that approach. To be consistent, the additional data masked by the relevance mask should also stick to the same method. Nevertheless, the insertion of random noise or mean values would also be worth testing to see how this might interfere with the predictions of the CNN or interrupts the decision process of the CNN classifier when trying to rely on spatial patterns. However, this evaluation can be seen as a simple test whether the CNN classifiers' decisions really rely strongly on the geographical regions obtained by the LRP algorithm. The basic idea is to compare the CNN's skill to random occluded data, meaning that the CNN classifier predicts on data where random grid cells were occluded before prediction. This can be seen as a basic idea for a metric evaluating the importance of the regions obtained through applying the LRP algorithm to the trained CNN classifier. As implemented by libraries as quantus (Hedström et al. 2023), this can be done quantitavely as well. Nevertheless, in this work several predictions were done by the CNN classifier, occluding the data step by step until an occlusion fraction of 10% with respect to the total amount of grid cells was reached. More precisely, 5 predictions with two percent steps with respect to the occluded grid cells were done for each the random occluded and the relevance based occluded data. The relevance based occlusion of the grid cells was chosen according to the highest relevance scores in magnitude (positive and negative scores) obtained by the explanations. For example, with an occlusion percentage of 2%, 1% of the highest scoring grid cells and 1% of the lowest scoring grid cells are occluded based on their relevance scores.(see 3.1. These grid cells were determined using the average explanation maps for each class as template.

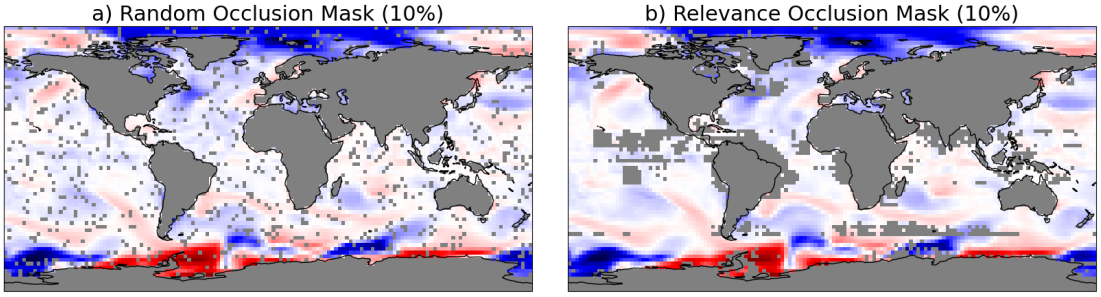


Figure 12: *Examples for the masked temperature maps used for evaluating the CNN classifier. For a) the grid cells were randomly occluded while in b) 10% of the most relevant cells are occluded. For better contrast the colormap 'seismic' was chosen in combination with a gray backround color. For this relevance mask, the explanations of the ABS data are taken as template, see Fig. 5.*

The usual metrics applied in the context of the evaluation of explanations as in Quantus (Hedström et al. Hedström et al. 2023) rely on the same approach. Methods such as region perturbation are already implemented in such frameworks, but destined for 3D data. Therefore, the region perturbation was implemented as explained above. For a more quantitative approach, it is also possible to iteratively occlude every pixel and measure the difference in the accuracy. However, following the strategy of iteratively occluding a fraction of the temperature maps before prediction, one can find that the accuracy drops significantly (see Fig.13). While in the DES case an occlusion of 10 percent only leads to an accuracy drop of a few percent in the random case, the accuracy through the relevance based occlusion drops down by over 18%. These results suggest that the geographical regions obtained by applying the LRP algorithm indeed play a decisive role for a correct prediction. Therefore, some underlying distinct biases have to exist within these regions that the CNN is capable of capturing.

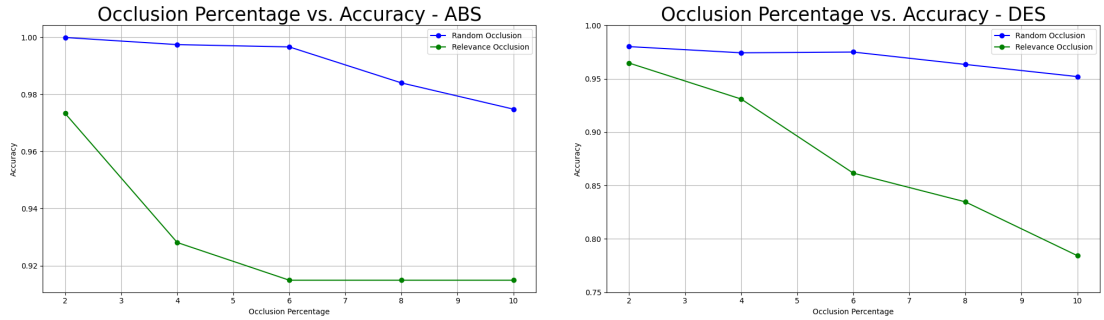


Figure 13: *Accuracy decrease upon increasing occlusion for the two differently preprocessed datasets.*

Regarding the accuracy drop in the ABS case (Fig.13), the experiment revealed that the accuracy quickly approaches a plateau that doesn't change upon increasing occlusion of the data. This is due to the fact, that the classifier is identifying all the samples as climate model output. The reason for the high absolute accuracy is that much more climate model data was used for the prediction step. Already for an occlusion fraction of 4%, the CNN classifier does only identify a few OBS samples correctly. Regarding that the CNN classifier for the ABS case has the highest accuracy, this is an interesting observation. For an occlusion fraction of 6%, the classifier does not identify any OBS samples anymore as OBS. They are all predicted as MOD.

Considering the case of the DES data, the accuracy drops much more because the wrong predictions for the MOD increases substantially as well. Compared to the ABS data, the wrong predictions distribute more evenly over the different datasets. This leads to a more severe drop in the absolute accuracy for this experiment, nevertheless the classifier still identifies samples for both classes, while this is not the case for the CNN classifier trained on the ABS data. For some climate models of the MOD class in the DES case, the wrong predictions increase very fast compared to the OBS. While the vast majority of the OBS is still correctly identified for an occlusion fraction of 10%, the wrong predictions even exceed 50% for one climate model, while the wrong predictions for other climate models occur much more often compared to the OBS as well (see Fig.14).

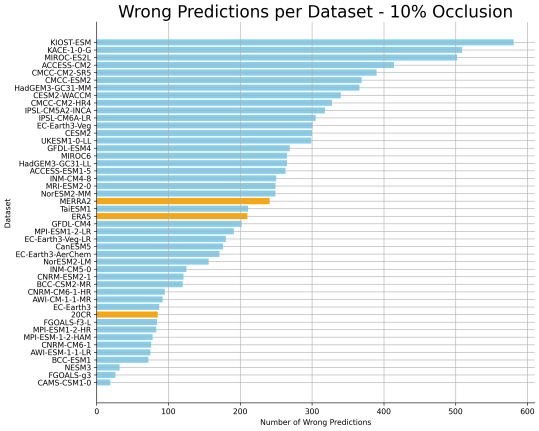


Figure 14: *Example count for predictions of the datasets that were occluded with respect to the relevance scores, in the DES case. In contrast to the wrong predictions for the ABS data, they are not that heavily imbalanced in this case, even leading to more misclassifications among the MOD.*

It is probably worth mentioning, that the DOISST dataset is never misclassified. As already seen in the data dependent explanations of the OBS, those for DOISST obviously differed from the other three reanalysis datasets. Since the relevance occlusion is based on the average of the explanations for the entire classes (MOD and OBS) the DOISST specific differences are probably not reflected enough in the relevance mask used for occlusion. Therefore, it could be that the occluded DOISST data still exhibits DOISST specific differences that are easily accessible for the CNN classifier.

4.2.7 Occluded Training

Having applied the occlusion procedure to already trained CNN classifiers, the question arose how this approach would affect the CNN classifiers performance if applied before training. For this experiment, 20% of the most relevant grid cells were occluded before training, resulting in much less (relevant) input fetaures. Nevertheless it was interesting, how the CNN would handle this scenario and how it would be able to still extract robust differences between the OBS and MOD.

In the following, only the results for the CNN classifier trained on the DES data are provided, since upon the end of this project the author was lacking computational resources due to turbulences at the server used for all the computations.

Contrary to the findings in 4.2.6, where half of the MOD samples were more often wrongly predicted than all the OBS, the opposite trend seems to be the case when occluding the samples before the training step. However, it is important to emphasize the high occlusion fraction of 20% compared to the 10% for the experiment where the already trained CNN classifier predicted the occluded data.
Within this experiment, the number of wrongly predicted OBS was always higher. Nevertheless, the gap to MOD that were misclassified the most was not that severe (see Fig 13). However, the samples of MOD were much more often correctly identified than the OBS. The worst accuracy regarding the OBS was exhibited for MERRA2, for which it

dropped to around 60% for one of the experiments. Also, for the first time, DOISST samples were among the wrongly predicted samples, but still very rarely.

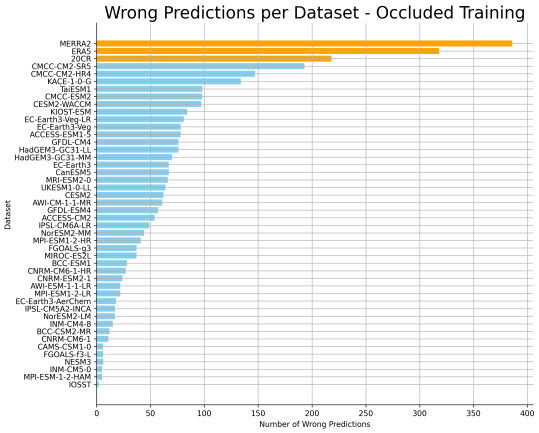


Figure 15: *Example for one of the conducted experiments. Prediction of CNN classifier trained on 20% fraction occluded data. Wrong predictions are biased towards the OBS.*

Again, it shall be mentioned that the number of conducted prediction experiments (n=5) for the corresponding classifier is quite low, due to the fact that the variations among the predictions were always found to be marginal. Nevertheless, the author wants to point out that the CNN classifiers' variance within its predictions was not thouroughly tested, because it requieres computational effort despite being immensely time consuming. However, throughout all the experiments, the trained CNN classifiers were observed to be stable in terms of their predictions.

Anyway, the occlusion before the training step again shows a significant decrease in accuracy. But still, even for occluding such a large fraction of the most relevant grid cells, the CNN classifier is able to establish some skill. It is definitely not an accurate CNN classifier anymore, but the skill is not neglectable, since on average it still correctly identifies the vast majority of samples for both classes. However, this experiment shall not just provide explanations of a less accurate CNN classifier.

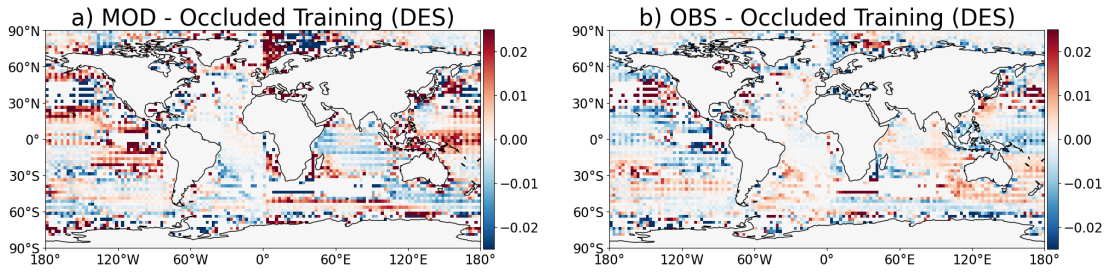


Figure 16: *Explanations for the CNN trained on occluded data, DES case.*

Instead, this should be viewed as a final experiment within the scope of this project, addressing the challenge of estimating the amount of information a CNN can extract from specific geographical regions. Of course it is straight forward to apply certain metrics and

evaluate on the reliability of a given CNN classifier. But nevertheless, we are far away from having a real insight into the decision making process using these methods. A first attempt that appears tempting would be to narrow the amount of input features down to the most important ones to see the effect in the accuracy. It could be applied as a filtering process, evaluating the amount of information that lies within certain geographical regions. This approach of decomposing the feature space before the training process eventually reveals an estimator for the amount of information that can be expected for e.g. the comparison between OBS and MOD. Or to put it differently, it may reveal the necessary amount of grid cells for a robust distinction between OBS and MOD.
However, when looking at the explanations for a CNN classifier trained on occluded data, it appears quite messy - almost looking like noise. Missing many highly relevant geographical regions, there are no clear geographical patterns observable anymore. It seems as if the neighboring grid cells of the important regions still contain valuable information (see Fig 16). Mostly, the highest scores are surrounding the important regions that have been removed. It would be for instance an interesting question, at which point of grid cell removal, the skill completely vanishes or drops below a certain level of e.g. an accuracy of 50%. This may also reveal important spatial dependencies between grid cells. To find this out, one would need to iteratively occlude relevant areas before training, evaluate the predictions and repeat this process until a specified level of accuracy has been reached. Removing relevant areas to measure the effect in the accuracy of the CNN classifiers predictions could help in estimating the amount of underlying information (biases) within specific regions. It could serve as a first orientation where to look at more precisely.
Finally, it boils down to the question, how many input features are necessary for a skillful prediction. In this context, that could be translated into the question, what geographical regions deliver most of the necessary information for this distinction between MOD and OBS, which is the basic approach during climate model evaluation. By looking at isolated regions, it may be possible to determine their levels of information content which the CNN classifier is using for the accurate predictions.

It shall be pointed out that the relevance scores obtained by such XAI methods are always relative scores. Moreover, as shown in this work, they are heavily dependent on the provided training samples. This is a crucial fact since we don't know how these relevance scores can be taken e.g. upon occlusion. How do they change if occluding a certain area and why? Is this change in relevance really worth investigating? Or upon the removal of the mean seasonal cycle, as it was observed in 4.2.2, where regions as e.g. the indian ocean flipped the sign in the relevance scores from positive to negative and vice versa, depending on the class (observational data or climate model output). Such changes are not straightforward to understand. These complex spatial dependencies between certain geographical regions or single grid cells are hard to disclose, the results of LRP don't deliver the reason why the regions are considered to be highly relevant.

To conclude the results section, it shall be emphasized that all these experiments used the average of all the generated explanations per temperature map. Indeed, explanations for individual datasets were provided to show important differences among the datasets (4.2.3), but it is probably worth it considering a closer look into the indivual explanations. The plots of the individual climate model explanations are provided in the appendix (7). Also, it could be interesting for instance in the ABS case, to look at the different seasons

of the year and how the explanations change.

However, there are ample opportunities for using the results of such explanations obtained by e.g. LRP. It is a first step towards understanding the decision-making process of a CNN. But, to get near the reasoning for the obtained relevance scores, much more has to be done.

5 Conclusions

This work shows the successful application of layerwise relevance propagation used for a binary CNN classifier that was trained to distinguish between observational based samples and climate model output. The samples during the training process are represented by daily temperature maps. The grid cells of these temperature maps served as the input features the CNN learned from, which can be seen as pixels of a digital image. Moreover, each class consisted of various datasets: 43 climate models were used and 4 observational datasets. Through the procedure of propagating backwards from the output layer to the input layer using laywerwise relevance propagation (LRP), it is possible to obtain heatmaps or explanations as referred to in this work. These are showing which input features of the daily temperature maps contribute the most to the classifiers decision.

1. **Which geographical regions are most important for a skillful prediction with respect to daily temperature maps?**

The resulting heatmaps immediately provide the geographical regions that are most influencial in terms of the binary CNNs' decisions. Depending on the different preprocessing approaches (removal of global mean only or mean seasonal cycle in addition), these geographical regions differ to some extent. Nevertheless, when taking a closer look, many highly relevant areas persist for both data types. These would be e.g. the North Atlantic, the Southern Ocean, Northern South America, the Equatorial Pacific as well as the Indican Ocean to name the most salient regions. Interestingly, a highly important region at the transition from western southern-Africa to central Africa loses its relevance upon the removal of the mean seasonal cycle. Additionally it could be observed that positive contributions happen to flip to negative contributions upon the removal of the mean seasonal cycle. Such dynamics are not trivial to understand and require further investigations of the specific datasets used or the specific geographical regions. For instance, the impact on the explanations, which reflect the relevance scores of the input features, was quite remarkable when one of the four observational datasets was removed. As shown, the results differ considerably, but still exhibiting geographical regions such as the Southern Ocean, western southern-Africa, equatorial Pacific to be very relevant. However, through the computation of the standard deviation, using the explanations of both classes, it was aimed to get an overview of the most relevant regions. While this applied to the case of the ABS data, it did not perfectly work in the DES case. The relevance scores are inversely related and should exhibit larger standard deviations for regions to which high relevance scores in magnitude where assigned. As found during the experiments, the explanations in the DES case did not exhibit this strong inverse pattern, therefore different areas for each class accumulated higher relevance scores, leading to less overlap between explanations of the observational data and climate model output. Therefore, areas that are highly important to only one class, don't show such a high standard deviation, since they were not equally weighted (with the opposite sign) for the other class. However, it is an interesting observation how the CNN distributes the relevance scores in a different way with respect to each class it predicts in the DES case.
Evaluation of the explanations provided by LRP showed that occluding relevant areas before the classifier made predictions significantly impacted the accuracy, highlighting the importance of these regions for the CNN classifier's performance.

2. **To what extent is it possible to pinpoint the predictive skill of CNNs when identifying climate models?**

It is definetely a very complex task to gain insights into the decision-making process of a CNN classifier. However, additional application of rather classical statistics can reveal some insights into the classifiers behaviour. Through looking at the correlations between the explanations of the two differently preprocessed datasets, it could be shown that the correlation between the explanations significantly drops after removal of the mean seasonal cycle. This would support the assumption, that upon the removal of information or biases, the classifier is forced to learn from more complex spatial patterns in the data. Since during the learning process a specific loss function is iteratively minimized, the CNN will always learn as much as necessary to fulfill a high accuracy hence a minimization of the loss function.

While the application of LRP to the misclassified samples just resulted in the same relevance scores (just assigned to the wrong class of datasets), a closer look at the number of misclassifications gave more insights. For instance, when only reanalysis datasets where used during training, MERRA2 and ERA5 made up more than 50% of the wrong predictions. Including the DOISST dataset, which is very different from the other three reanalysis datasets, led to an decrease in wrong predictions for MERRA2 and ERA5 as well. This indicates that the CNN classifiers' accuracy could depend on the complexity of the feature space of each class. Having less variance in the feature space would lead to a worse accuracy in this setting. Moreover, these results indicate that there are underlying patterns the CNN classifier learns, that are shared between the reanalysis datasets and the climate models. The explanations for the reanalysis datasets were pretty homogenous while those for DOISST differed considerably in the DES case. The occlusion experiment revealed something interesting regarding the differently preprocessed datasets: while in the case where only the global mean was removed, the CNN already identified every sample as climate model output when only 6% of the grid cells were occluded. The decrease in accuracy for the DES case happened to be more steady. The reason for this is, that the wrong predictions distributed more evenly across all the different datasets, except for DOISST which had still not been wrongly predicted. However, regarding the fact that the accuracy seems to be more fragile in the ABS case, it seems as if the CNN classifier is able to learn more obvious differences between the classes. Regarding that the relevance scores obtained in this case shared a strong positive Pearson correlation, it would make sense that the CNN classifier is relying on similar features that are very dependent on each other. In the other case, this seems not to be the case at all, suggesting a more spatially complex learning behaviour of the CNN in the DES setting. The single relevance scores don't share any considerable correlation anymore. Looking at the explanation maps, it is also very obvious how the CNN more evenly assigns scores to the grid cells, showing less concentrated areas of high relevance.

Through the occlusion of the samples before using them for training, it could be shown that such results obtained by a CNN can indeed be counter intuitive. During the occlusion of the already trained CNN classifier, it was observed that the CNNs' accuracy drops more drastically for the ABS data. However, the accuracy for the other DES data drops more steadily but also deeper, since the CNN starts to wrongly classify the samples in both directions: it doesn't, as in the ABS case, always identify all the samples as climate model output, it confuses both classes. Also, the highest counts for wrongly

predicted samples belonged to the climate model outputs.After occluding the samples before training though, the results showed that all the reanalysis datasets exhibit the highest numbers of misclassifications. Also, the DOISST data appears for the first time among the wrong predictions, but only a few samples could be found to be misclassified during the experiments.

Regarding such dynamics, it turns out that the question, how to exactly use XAI methods to support this domain, are far from trivial. The provided approaches to occlude e.g. daily temperature maps and test how far valuable information can still be extracted, are valid. It can be used to quickly identify important regions that play a decisive role when comparing observational data to climate model output, as it is done for climate model evaluation. However, applying such methods, we will again have to face a very hard problem: finding out why the CNN assigned a certain relevance to a specific grid cell. Again, we are confronted with a result, just from the other side of the network: we start with having an accurate classifier delivering correct predictions, then apply methods as LRP and again obtain results lacking any reasoning. For tasks as climate model evaluation though, which require a very detailed knowledge of potential error sources, that may not be sufficient. For this reason, it is really important to develop methods that not only allow for a reliable evaluation of the explanations, but also provide more information on the learning-behaviour of such a network. Under certain circumstances, it may be enough to know what e.g. pixels or general input features are most important to conclude on the reason why they are so relevant. But regarding the task of climate model evaluation, there are too many different datasets involved, all containing their own biases that won't be exactly the same. For being able to get to the exact patterns the CNN learnd e.g. during the experiments in this work, much more investigations would be necessary to achieve that. It is definitely a huge progress having access to the relevance of every single input feature compared to the others, but if the reasoning why certain regions are important is lacking, it will be hard to get to the root cause why a CNN can learn robust differences between observational data and climate model data already at a daily scale.

6 Outlook

It will be an increasingly hot topic to get more insights into the decisions of ANN's in general. This comprises all the domains where applied. It could be incredibly useful for the evaluation of climate models to gain more insights into a CNN that was properly trained to find distinctive patterns within the observational based datada and climate model output. Therefore, this work can be expanded in numerous directions, starting from the evaluation of such a classifier to the extraction of specific geographical areas for examination. The examination of such a CNN would require to dive deeply into the details of the architecture as well. So, any future work continuing the application of machine learining in climate science or the context of climate model evaluation would need an interdisciplinary team, since the specific domain knowledge as well as the technical knowledge are immensely important.

Despite of focusing on the decision-making process of such a CNN, there are many potential evaluation tasks worth investigating. For instance the influence of certain datasets on the predictions or overall accuracy of a CNN. As observed in this work, the use of the DOISST dataset turned out to have a relatively strong impact on the classifiers accuracy. Examining how the bias in the training data leads to a bias in the predictions of a CNN would be important to fully understand the reasons for the explanations obtained by e.g. LRP.

Another idea that can be extented in numerous ways is to look at the different geographical regions received by an explanation in more detail. By decomposing the training data in accordance to the explanations, it may be possible to encounter underlying patterns that shed light on the different underlying biases the CNN classifier learns from. Moreover, applying LRP to a multiclass CNN classifier trained to identify individual datasets, rather than just distinguishing between observational data and climate model output, can reveal more dataset-specific differences.

In general, it is definitely a reasonable approach to use XAI techniques in settings where the predicted classes are not that broadly defined as in the case of this binary CNN classifier. For climate model evaluation, it could be useful to train the classifier on subsets of climate models that shall be compared or tested. As also shown by Brunner and Sippel (2023), the multiclass CNN classifier wrongly predicted expecially those climate model outputs that were generated by the same model families developed at the same institutes, suggesting dependencies within the climate models through e.g. shared code. Such findings could be further tested by setting up a CNN that was trained e.g. on specific model families. By applying techniques as LRP to such a classifier, this could maybe reveal geographical areas where the climate model families differ the most. The areas found could then be investigated in detail.

Of course, the same could be applied to observational data. Given that observational uncertainties are challenging to detect but crucial for accurate climate model evaluation, combining machine learning with explainable AI (XAI) may prove valuable in identifying tendencies or biases within the various observational datasets. This could help to improve the understanding of observational uncertainty.

7 Appendix

Libraries Used

- **Python**: 3.9.0
- **TensorFlow**: 2.14.0
- **Innvestigate**: 2.0.1
- **NumPy**: 1.25.2
- **Cartopy**: 0.22.0
- **Xarray**: 2023.10.1
- **Matplotlib**: 3.8.0

Operating System

OS Version:

```
Linux 4.18.0-477.10.1.el8_8.x86_64
Build Information: #1 SMP Tue May 16 11:38:37 UTC 2023
Detailed OS Info: Linux-4.18.0-477.10.1.el8_8.x86_64-x86_64-with-glibc2.28
```

CNN Architecture of Deep Classifier

Layer (type)	**Output Shape**	**Param #**
conv2d_4 (Conv2D)	(None, 36, 72, 128)	3328
max_pooling2d_4 (MaxPooling2D)	(None, 9, 18, 128)	0
conv2d_5 (Conv2D)	(None, 9, 18, 64)	73792
max_pooling2d_5 (MaxPooling2D)	(None, 4, 9, 64)	0
conv2d_6 (Conv2D)	(None, 4, 9, 64)	36928
max_pooling2d_6 (MaxPooling2D)	(None, 2, 4, 64)	0
conv2d_7 (Conv2D)	(None, 2, 4, 32)	8224
max_pooling2d_7 (MaxPooling2D)	(None, 1, 2, 32)	0
flatten_1 (Flatten)	(None, 64)	0
dense_1 (Dense)	(None, 2)	130
Total params		**122402 (478.13 KB)**
Trainable params		122402 (478.13 KB)
Non-trainable params		0 (0.00 Byte)

CNN Architecture of the Simplified Model

Layer (type)	**Output Shape**	**Param #**
conv2d (Conv2D)	(None, 18, 36, 64)	640
max_pooling2d (MaxPooling2D)	(None, 9, 18, 64)	0
conv2d_1 (Conv2D)	(None, 9, 18, 128)	73856
max_pooling2d_1 (MaxPooling2D)	(None, 4, 9, 128)	0
conv2d_2 (Conv2D)	(None, 4, 9, 256)	295168
max_pooling2d_2 (MaxPooling2D)	(None, 2, 4, 256)	0
flatten (Flatten)	(None, 2048)	0
dense (Dense)	(None, 2)	4098
Total params		**373762 (1.43 MB)**
Trainable params		373762 (1.43 MB)
Non-trainable params		0 (0.00 Byte)

Climate Models used: Simplified CNN Classifier

- ACCESS-CM2
- AWI-CM-1-1-MR
- BCC-CSM2-MR
- CAMS-CSM1-0
- CESM2
- CMCC-CM2-HR4
- CanESM5
- EC-Earth3-AerChem
- EC-Earth3-Veg-LR
- EC-Earth3-Veg
- EC-Earth3
- HadGEM3-GC31-MM
- INM-CM5-0
- IPSL-CM6A-LR
- KACE-1-0-G
- KIOST-ESM
- MIROC6
- MPI-ESM1-2-LR

- MRI-ESM2-0
- NESM3
- NorESM2-LM
- NorESM2-MM
- TaiESM1
- UKESM1-0-LL

Observational Datasets used: Simplified CNN Classifier

- 20CR
- ERA5
- MERRA2

Climate Models used: Deep CNN Classifier

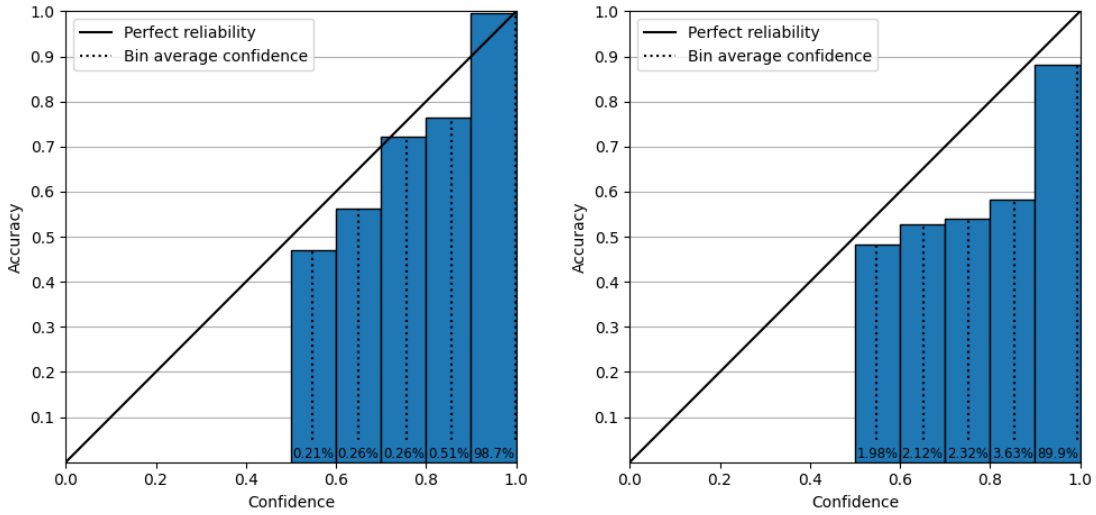
- ACCESS-CM2
- ACCESS-ESM1-5
- AWI-CM-1-1-MR
- AWI-ESM-1-1-LR
- BCC-CSM2-MR
- BCC-ESM1
- CAMS-CSM1-0
- CanESM5
- CESM2
- CESM2-WACCM
- CMCC-CM2-HR4
- CMCC-CM2-SR5
- CMCC-ESM2
- CNRM-CM6-1
- CNRM-CM6-1-HR
- CNRM-ESM2-1
- EC-Earth3

- EC-Earth3-AerChem
- EC-Earth3-Veg
- EC-Earth3-Veg-LR
- FGOALS-f3-L
- FGOALS-g3
- GFDL-CM4
- GFDL-ESM4
- HadGEM3-GC31-LL
- HadGEM3-GC31-MM
- INM-CM4-8
- INM-CM5-0
- IPSL-CM5A2-INCA
- IPSL-CM6A-LR
- KACE-1-0-G
- KIOST-ESM
- MIROC-ES2L
- MIROC6
- MPI-ESM-1-2-HAM
- MPI-ESM1-2-HR
- MPI-ESM1-2-LR
- MRI-ESM2-0
- NESM3
- NorESM2-LM
- NorESM2-MM
- TaiESM1
- UKESM1-0-LL

Observational Datasets used: Deep CNN Classifier

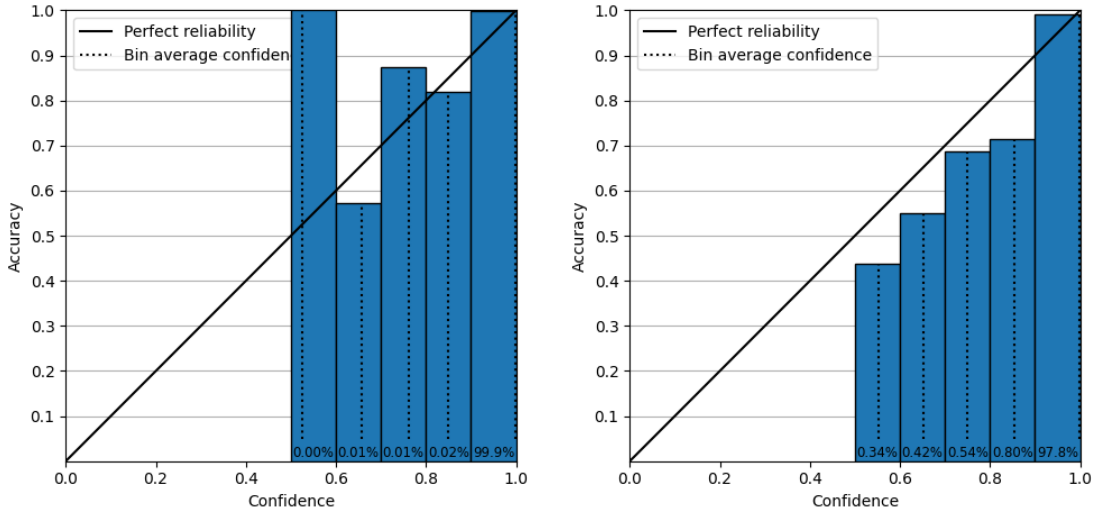
- 20CR
- ERA5
- DOSST
- MERRA2

Reliability Diagrams for the Simplified Model



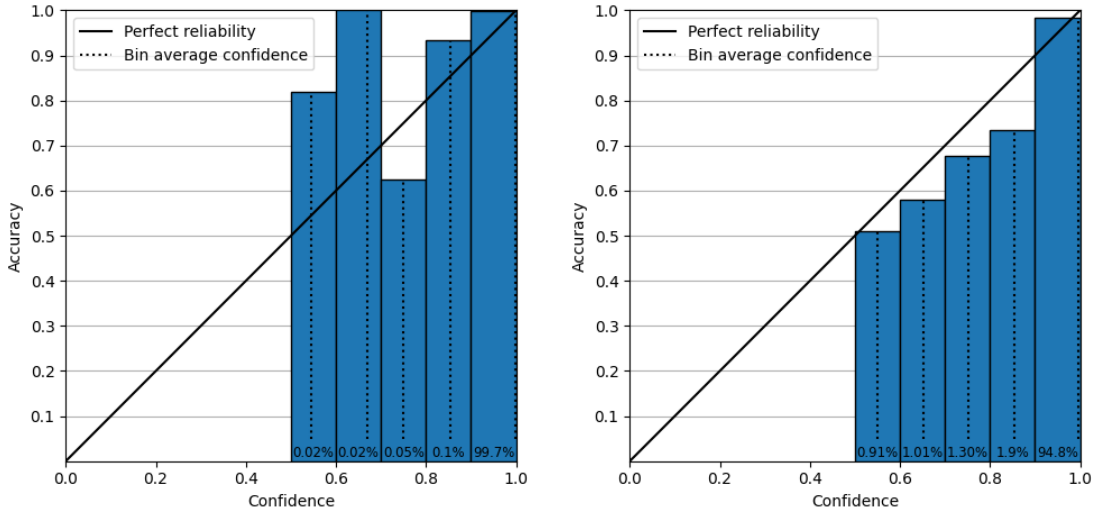
Reliability plots for the toy model in the ABS case on the left, deseasonlaized case on the right hand side.

Reliability Diagrams for the Deep Classifier - with DOISST



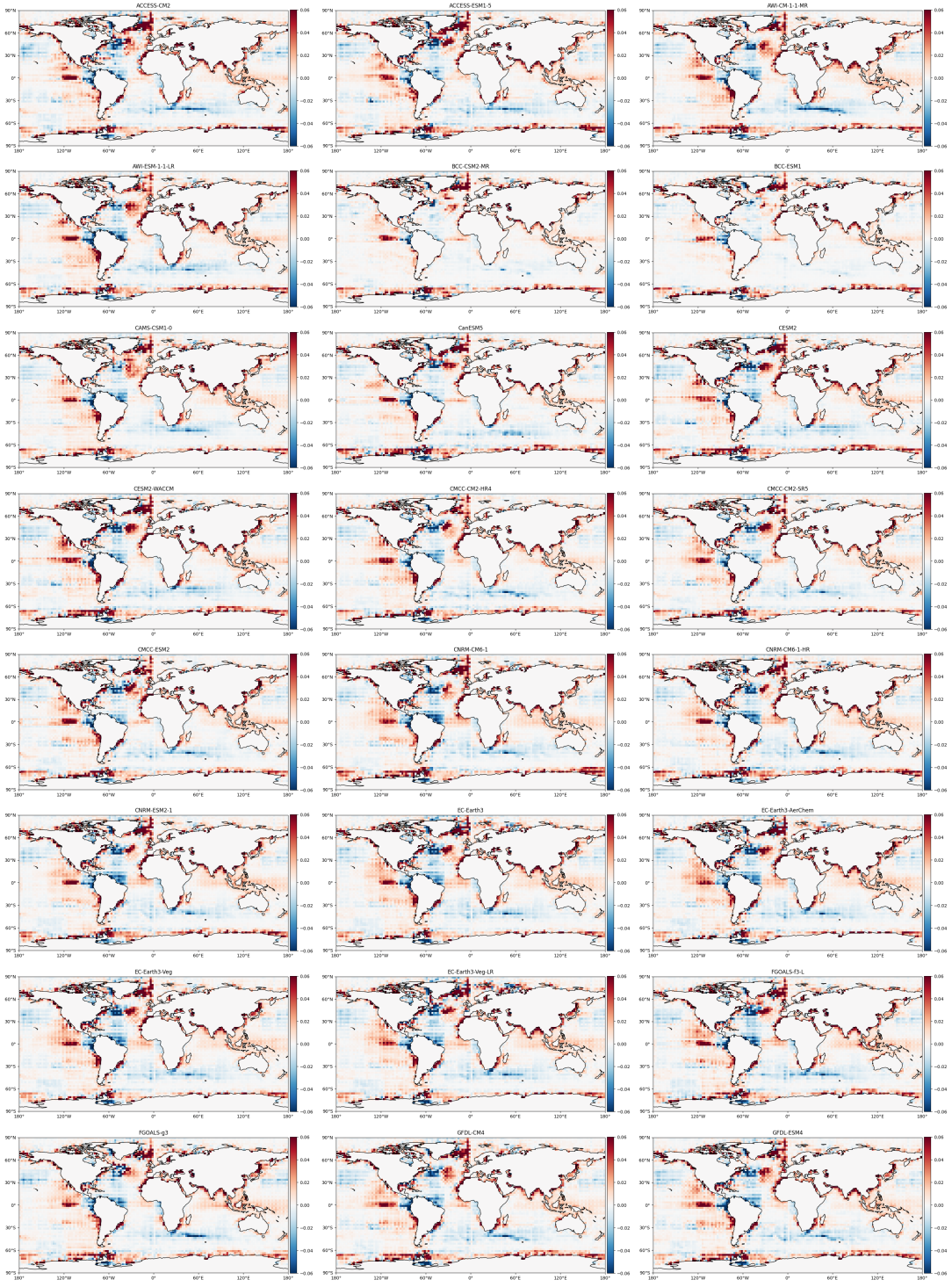
Reliability plots for deep classifer trained with DOISST data as well. In the ABS case on the left, deseasonlaized case on the right hand side.

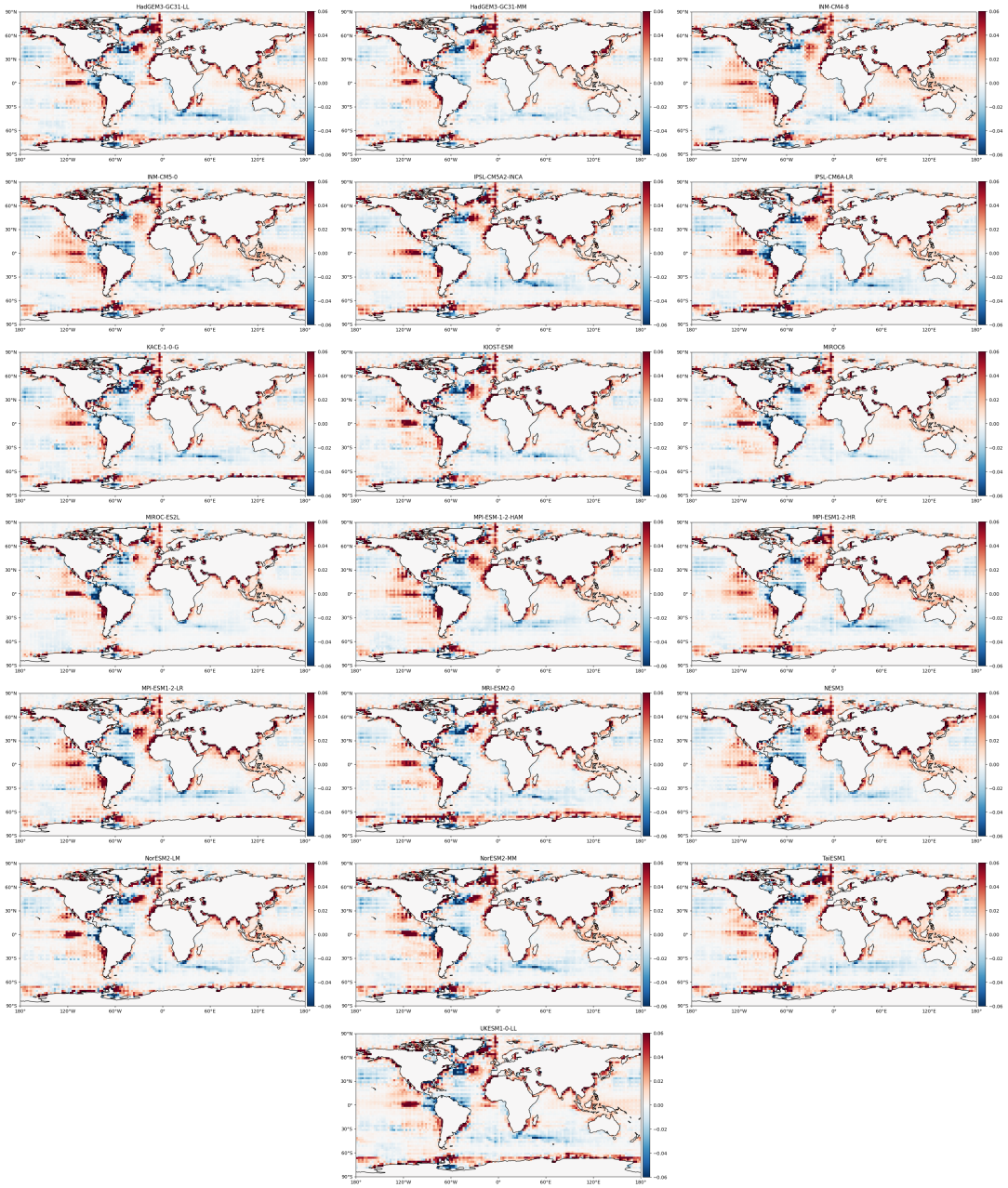
Reliability Diagrams for the Deep Classifier - without DOISST



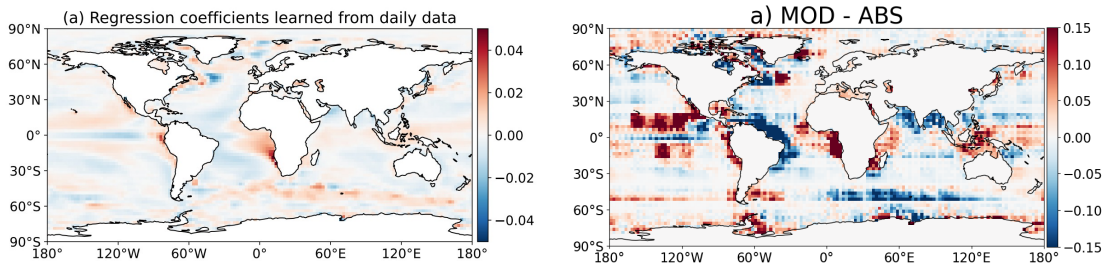
Reliability plots for deep classifer trained without DOISST data. In the ABS case on the left, deseasonlaized case on the right hand side.

Average Explanation per Climate Model




HadGEM3-GC31-LL
HadGEM3-GC31-MM
INM-CM4-8
INM-CM5-0
IPSL-CM5A2-INCA
IPSL-CM6A-LR
KACE-1-0-G
KIOST-ESM
MIROC6
MIROC-ES2L
MPI-ESM-1-2-HAM
MPI-ESM1-2-HR
MPI-ESM1-2-LR
MRI-ESM2-0
NESM3
NorESM2-LM
NorESM2-MM
TaiESM1
UKESM1-0-LL

Comparison with Regression Coefficients by Brunner and Sippel (2023)



Direct comparison between the explanations of the binary CNN to the obtained coefficients of a simplper linear regression model produced by Brunner and Sippel (2023).

References

Alber, Maximilian et al. (2019). "iNNvestigate Neural Networks!" In: *Journal of Machine Learinng Research*. DOI: https://doi.org/10.48550/arXiv.1808.04260.

Annan, James D. and Julia C. Hargreaves (2017). "On the meaning of independence in climate science". In: *Earth System Dynamics, 8, 211–224*. DOI: doi:10.5194/esd-8-211-2017.

Baehrens, David et al. (2010). "How to Explain Individual Classification Decisions". In: *Journal of Machine Learning Research*. DOI: https://doi.org/10.48550/arXiv.0912.1128.

Bommer, Philine Lou et al. (2024). "Finding the Right XAI Method—A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science". In: *American Meteorological Society*. DOI: DOI:10.1175/AIES-D-23-0074.1.

Brunner, Lukas and Sebastian Sippel (2023). "Identifying climate models based on their daily output using machine learning". In: *Cambridge University Press*. DOI: https://doi.org/10.1017/eds.2023.23.

CarbonBrief (2018). *Q & A: How do climate models work?* URL: https://www.carbonbrief.org/qa-how-do-climate-models-work/.

Eyring, Veronika et al. (Feb. 2019). "Taking climate model evaluation to the next level". In: *Nature Climate Change* 9, pp. 102–110.

Gettelman, Andrew and Richard B. Rood (2016). *Demystifying Climate Models, Version 2*. Springer.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. Draft of August 10, 2016. MIT Press. URL: libgen.li/file.php?md5=e4b2ab0ef22458f94c835d4d2397034e.

Hedström, Anna et al. (2023). "Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond". In: *Journal of Machine Learinng Research*. DOI: https://doi.org/10.48550/arXiv.2202.06861.

Huang, Boyin et al. (2020). "Improvements of the Daily Optimum Interpolation Sea Surface Temperature (DOISST) Version 2.1". In: *American Meteorological Society*, pp. 2923–2939. DOI: DOI:10.1175/JCLI-D-20-0166.1.

Iturbide, M. et al. (2020). "An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets". In: *Earth System Science Data* 12.4, pp. 2959–2970. DOI: 10.5194/essd-12-2959-2020. URL: https://essd.copernicus.org/articles/12/2959/2020/.

J., Berman Jules (2016). *Data simplification : taming information with open source tools*. 1st ed. Morgan Kaufmann (imprint of Elsevier).

Merrifield, Anna L. et al. (2023). "Climate model Selection by Independence, Performance, and Spread (ClimSIPS v1.0.1) for regional applications". In: *Geoscientific Model Development, 16, 4715–4747*. DOI: https://doi.org/10.5194/gmd-16-4715-2023.

Montavon, Gregoire et al. (2019). "Layer-Wise Relevance Propagation: An Overview". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 193–209. DOI: DOI:10.1007/978-3-030-28954-6_10.

O'Shea, Keiron and Ryan Nash (2015). "An Introduction to Convolutional Neural Networks". In: *arXiv*. DOI: https://doi.org/10.48550/arXiv.1511.08458.

Parker, Wendy S. (2015). "REANALYSES AND OBSERVATIONS - What's the Difference?" In: *American Meteorological Society*. DOI: DOI:10.1175/BAMS-D-14-00226.1.

Samek, Wojciech and Klaus-Robert Müller (2019). "Towards Explainable Artificial Intelligence". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, vol. 11700, pp. 5-22.* DOI: https://doi.org/10.1007/978-3-030-28954-6_1.

Sippel, Sebastian, Nicolai Meinshausenvand Erich M. Fischer Enikő Székely, and Reto Knutti (2020). "Climate change now detectable from any single day of weather at global scale". In: *Nat. Clim. Chang. 10, 35–41.* DOI: https://doi.org/10.1038/s41558-019-0666-7.

Zumwald, Marius et al. (2019). "Understanding and assessing uncertainty of observational climate datasets for model evaluation using ensembles". In: *WIREs Climate Change published by Wiley Periodicals LLC.* DOI: DOI:10.1002/wcc.654.