A data-driven probabilistic network approach to assess model similarity in CMIP ensembles

Catharina Elisabeth Graafland, a Swen Brands, a and José Manuel Gutiérrez a

a Instituto de Física de Cantabria, CSIC–Universidad de Cantabria, E-39005, Santander, Spain

Corresponding author: Catharina Elisabeth Graafland, catharina.graafland@unican.es

Early Online Release: This preliminary version has been accepted for publication in Artificial Intelligence for the Earth Systems, may be fully cited, and has been assigned DOI 10.1175/AIES-D-23-0073.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

© 2024 American Meteorological Society. This is an Author Accepted Manuscript distributed under the terms of the default AMS reuse license. For information regarding reuse and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
ABSTRACT: The different phases of the Coupled Model Intercomparison Project (CMIP) provide ensembles of past, present, and future climate simulations crucial for climate change impact and adaptation activities. These ensembles are produced using multiple Global Climate Models (GCMs) from different modeling centres with some shared building blocks and inter-dependencies. Applications typically follow the ‘model democracy’ approach which might have significant implications in the resulting products (e.g. large bias and low spread). Thus, quantifying model similarity within ensembles is crucial for interpreting model agreement and multi-model uncertainty in climate change studies. The classical methods used for assessing GCM similarity can be classified into two groups. The *a priori* approach relies on expert knowledge about the components of these models, while the *a posteriori* approach seeks similarity in the GCMs’ output variables and is thus data-driven. In this study we apply Probabilistic Network Models (PNMs), a well established machine learning technique, as a new *a posteriori* method to measure inter-model similarities. The proposed methodology is applied to surface temperature fields of the historical experiments from the CMIP5 multi-model ensemble and different reanalysis gridded datasets. PNM s are capable to learn the complex spatial dependency structures present in climate data, including teleconnections operating on multiple spatial scales, characteristic of the underlying GCM. A distance metric building on the resulting PNM s is applied to characterize GCM model dependencies. The results of this approach are in line with those obtained with more traditional methods, but have further explanatory potential building on probabilistic model querying.
SIGNIFICANCE STATEMENT: The present study proposes the use of Probabilistic Network Models (PNMs) to quantify model similarity within ensembles of Global Climate Models (GCMs). This is crucial for interpreting model agreement and multi-model uncertainty in climate change studies. When applied to climate data (gridded global surface temperature in this study) PNMs encode the relevant spatial dependencies (local and remote connections). Similarities among the PNMs resulting from different GCMs can be quantified and are shown to capture similar GCM formulations reported in previous studies. Differently to other machine learning methods previously applied to this problem PNMs are fully explainable (allowing probabilistic querying) and are applicable to high-dimensional gridded raw data.

1. Introduction

The Coupled Model Intercomparison Project (CMIP) provides numerical simulations of the past and future (under different scenarios) temporal evolution of the Earth system from a large number of nominally different Global Climate Models (GCMs) (Taylor et al. 2012; Eyring et al. 2016). GCMs typically include the physical components of the climate system, comprising atmosphere, land-surface, ocean and sea-ice. In addition to the physical components, other sub-models can be included to take into account the effects of growing vegetation, aerosols, atmospheric chemistry, terrestrial and ocean carbon-cycle processes or land-ice dynamics in what is then referred to as an “Earth System Model” (Séférian et al. 2019; Jones 2020; Brands 2022a).

CMIP aims to compare and improve these models, fostering a better understanding of climate processes and enhancing the reliability of future climate projections. To achieve this goal, CMIP has undergone several phases, with CMIP6 (the sixth phase) being the latest and most recent. In each phase, more institutions have participated than in the previous one, typically contributing with one to three GCMs (up to nine in CMIP6). In the latest phases, GCMs often have multiple runs provided, incorporating slightly different initial conditions or variations in the model’s physical parameters and processes.

The output data from these experiments forms the basis of manifold downstream climate change impact and adaptation activities, involving virtually all socioeconomic sectors (e.g., energy, agriculture, and health). The metrics drawn therefrom, climatological mean values in the simplest case, typically follow the ‘model democracy’ approach (also referred to as ‘one model, one vote’),

3

Accepted for publication in Artificial Intelligence for the Earth Systems. DOI 10.1175/AIES-D-23-0073.1.
in which all GCMs are considered equally plausible (Masson and Knutti 2011a; Knutti et al. 2013, 2017). Ideally, this multi-model ensemble should comprise independent model formulations that agree well with observations for phenomena operating on scales resolved by the models. However, institutions continuously improve their GCMs by building on sub-components of former models and also share parts of current models with other institutions. A ‘democratic’ multimodel ensemble might tend towards the more popular models/blocks that could have been chosen for various reasons. For example, institutes might have confidence in their performance, or choices are made because of available code or development and support resources. The main issue with ‘popular’ models is the fact that they distort ensemble statistics and can lead to low spread and inappropriate conclusions about certainty and robustness. In addition, biases in popular models/blocks can propagate to their similar counterparts, thus contributing to a common model bias. These factors may have significant implications for the bias and spread of the metrics drawn from the democratic multi-model ensemble output.

To weight the GCMs according to their dependencies, two main approaches can be distinguished in the literature (Boé 2018). The first, “a priori” approach groups the models or assigns weights to them as a function of some kind of expert knowledge which, e.g., can be derived from a thorough metadata analysis. The simplest example is to put GCMs from the same institution(s) into the same group (Leduc et al. 2016; Annan and Hargreaves 2017). Going further, Boé (2018) grouped the GCMs according to the shared use of sub-models for the four classical components of a Climate System Model mentioned above. Brands (2022b) found similar spatial error patterns for those GCMs using Atmospheric Global Climate Model (AGCM) components of the same family, which was then confirmed by Merrifield et al. (2023) in an independent study, meaning that the predominant role of the AGCM for determining GCM similarities in the atmosphere is robust to changes in the specific experimental setup. Ideally, the a priori approach should be brought to a comprehensive source code analysis undertaken by the model developers themselves, but this is a heavily complex task that has only be accomplished for individual climate system components so far (Séférian et al. 2020). As an intermediate solution covering all components, Brands et al. (2023) have built an extensive metadata archive for more than 60 GCMs from CMIP5 and 6, containing the names and versions of up to 12 sub-models, resolution details, reference articles and other relevant information that can serve as a basis for further developing the a priori approach. Using
this source (Brands et al. 2023), it can be shown that the more than 60 GCMs used in CMIP5 and 6 rely on only 15 nominally different AGCMs, with some GCMs even using identical AGCM versions. The number of nominally independent sub-models for other components, e.g. the ocean, is even smaller.

The second approach to measure model dependencies assumes similarities in GCM output data (Abramowitz et al. 2019), usually climatological mean fields, or error fields with respect to observations, to be representative of model dependencies (Pennell and Reichler 2011; Bishop and Abramowitz 2013; Knutti et al. 2013; Boé 2018; Lorenz et al. 2018; Brunner et al. 2020). This is commonly done in terms of pairwise root mean square errors or Kullback-Leibler divergence (Boé 2018; Lorenz et al. 2018; Brunner et al. 2020; Knutti et al. 2013), but machine learning techniques are being applied more and more frequently, having the ability to learn nonlinear spatial relations between GCM output data. This approach does more justice to GCM outputs by capturing spatial and temporal non-linearities resulting from the systems of differential equations based on the fundamental laws of physics, fluid motion, and chemistry on which they are based. On the one hand, Brunner and Sippel (2023) use Convolutional Neural Networks (CNNs) to investigate whether models (and observations) have unique spatial features in their the output data that allow them to be identified even on daily timescales. On the other hand, Nowack et al. (2020) introduce causal model evaluation, an approach initially developed for climate data by Runge et al. (2014), as a type of process-oriented model evaluation. They quantify similarities between output data of climate models whose dimensions are reduced with varimax principal component analysis with a measure obtained from the networks describing the spatial features learnt from the data. The causal network learnt from a particular GCM provides a “fingerprint” of the global dynamics present in the dataset. In this context, Graafland et al. (2020) introduced Probabilistic Network Models (PNMs) to explore the most relevant spatial dependencies without the need for dimension reduction from climatological datasets, in this way capturing in the GCM output the interplay between local and global processes. The network topology and the associated probabilistic model reveal features of the underlying complex system that drove the dataset, and both can be analysed respectively with complex network or probabilistic measures to give insight in the spatial dependency structure of the climatological dataset. In this work we apply the PNM approach proposed in Graafland et al. (2020) to analyse the problem of model dependency. We use probabilistic networks to uncover the
spatial dependency structures within the historical experiments of a CMIP multi-model ensemble as well as two distinct reanalysis datasets. The differences in the learnt dependency structures are then used to estimate the inter-model dependencies within the ensemble. We show that probabilistic networks have the potential to fill the gap between the process-oriented but low dimensional causal networks and high dimensional CNNs that lack interpretation. A more in-depth discussion about the differences between CNNs, causal and probabilistic networks is provided in Section 4.

The paper is outlined as follows. In Section 2, we describe the applied datasets and introduce the basic concepts of Probabilistic Network Models, with a focus on Gaussian Bayesian Networks (GBNs), suitable for analyzing data describing complex systems. We also describe a probabilistic measure used to quantify the distance between two Bayesian Networks in a scalar, capable to take into account all features of the GBN backbone structure. In Section 3, we illustrate what the probabilistic and topological features of the obtained Bayesian Networks look like in an illustrative example ensemble comprising the historical experiments from four distinct CMIP5 models and one reanalysis. We show that the probabilistic distance measure captures the relevant features present in these datasets and then we apply it to the historical realizations of 25 additional GCMs and a second reanalysis. The main result is a powerful, direct and simple method to characterize these datasets according to their spatial dependency skeleton. Finally, in Section 4, we summarize the lessons learnt and compare our results with those obtained in previous studies relying on traditional and other machine learning methods.

2. Data and Methods

a. Reanalysis and CMIP5 Data

In this study, monthly-mean near-surface air temperature data from two reanalyses —ERA Interim (Dee et al. 2011) and JRA-55 (Kobayashi et al. 2015; Harada et al. 2016)— and from the historical experiments of 29 GCMs participating in CMIP5 (Taylor et al. 2012) are used on a global domain. The historical model runs were concatenated with the respective RCP8.5 scenario runs to cover the 30-year time period 1981-2010. The native resolution of the datasets varies from about 1° to 4°, but they were bilinearly interpolated to a 10° grid (approx. 1000 km), resulting in \( p = 648 \) grid points. Different resolutions were tested obtaining similar results and 10° was selected as a sensible compromise between sufficient resolution and computational efficiency.
Temperature anomaly values were obtained by removing the annual cycle (the 30-year mean values, month by month) from the raw data at each grid point $X_i$.

**b. Gaussian Bayesian Networks (GBNs)**

Spatial dependencies in climate data are the result of the interplay of manifold physical processes, resulting in both local and emergent distant dependence patterns, the latter commonly referred to as “teleconnections” (Rheinwalt et al. 2015). The configuration of climate models, including their components, parameterizations, coupling, etc., has an imprint on the modeling of the physical processes and consequently on the spatial dependencies in climate data. The aim of Probabilistic Network Models (PNMs) is to extract the backbone dependency structure, including both pairwise and high-order dependencies. PNMs are defined by a network topology (represented by a graph) and a probabilistic model (represented by the joint probability function) which can be learnt from data, revealing the structure of the underlying (complex) system.

In this work we use Gaussian Bayesian Networks (GBNs) as a subclass of PNMs to characterise the dependency structures in climatic gridded datasets (in particular reanalysis and GCM temperature data). Graafland and Gutiérrez (2022) show that this class of PNMs is most suitable for modelling high-dimensional data with a complex interaction structure. The term Gaussian refers to the choice of a multivariate Gaussian Joint Probability Density (JPD) function that associates graph edges with model parameters. The term Bayesian Network points to the type of parameters characterizing the JPD function and the way they are reflected by nodes and edges in the corresponding graph. The formulation and technical details of GBNs are explained in the next three paragraphs that parallel section ‘Probabilistic BN models’ in Graafland et al. (2020).

The multivariate Gaussian JPD function can take various representations in which dependencies between the variables are described by different types of parameters. The best-known representation of the Gaussian JPD function is in terms of marginal dependencies, *i.e.*, dependencies of the form $X_i, X_j | \emptyset$ as present in the covariance matrix $\Sigma$. Let $X$ be a $N$-dimensional multivariate Gaussian variable whose probability density function $P(X)$ is given by:

$$
P(X) = (2\pi)^{-N/2} \det(\Sigma)^{-1/2} \exp\{-1/2(X - \mu)^T \Sigma^{-1}(X - \mu)\},
$$

where $\mu$ is the $N$-dimensional mean vector and $\Sigma$ the $N \times N$ covariance matrix.
Alternatively, \( P(\mathbf{X}) \) in equation (1) can be characterized with conditional dependencies of the form \( X_i | S \) with \( S \subseteq \mathbf{X} \). The representation of the JPD is then a product of Conditional Probability Densities (CPDs):

\[
P(\mathbf{X}) = \prod_{i=1}^{N} P_i(\mathbf{X}_i | \Pi_{\mathbf{X}_i}) \tag{2}
\]

with

\[
P(\mathbf{X}_i | \Pi_{\mathbf{X}_i}) \sim \mathcal{N}(\mu_i + \sum_{j \in \mathbf{X}_j} \beta_{ij}(X_j - \mu_j), \nu_i) \tag{3}
\]

whenever the set of random variables \{\( X_i \mid \Pi_{\mathbf{X}_i} \)\}_{i \in N} is independent (Shachter and Kenley 1989). In this representation \( \mathcal{N} \) is the normal distribution, \( \mu_i \) is the unconditional mean of \( X_i \), \( \nu_i \) is the conditional variance of \( X_i \) given the set \( \Pi_{\mathbf{X}_i} \) and \( \beta_{ij} \) is the regression coefficient of \( X_j \), when \( X_i \) is regressed on \( \Pi_{\mathbf{X}_i} \). We call \( \Pi_{\mathbf{X}_i} \) the parentset of variable \( X_i \). In the context of climatological data \( \mathbf{X} \), imagine three variables \( X_i, X_j, \) and \( X_k \) representing temperature in three neighboring gridboxes, with \( X_j \) between the other two. Then, it could be that the correlation between \( X_i \) and \( X_k \) is fully explained by \( X_j \) and using the above notation this would render \( X_j \mid \Pi_{\mathbf{X}_j} \) independent of \( X_k \mid \Pi_{\mathbf{X}_k} \).

The probabilistic model of a Gaussian Bayesian Network is represented by equation (2). The corresponding graph of a GBN is a Directed Acyclic Graph (DAG) encoding the corresponding probability distribution as in equation (2). Each node corresponds to a variable \( X_i \in \mathbf{X} \) and the presence of an arc (i.e. connection) \( X_j \rightarrow X_i \) implies the presence of the factor \( P_i(X_i \mid \ldots X_j \ldots) \) in \( P(\mathbf{X}) \) and thus conditional dependence of \( X_i \) and \( X_j \). In this case, \( X_j \) is a parent of \( X_i \) and thus \( X_j \in \Pi_{\mathbf{X}_i} \). The absence of an arc between \( X_i \) and \( X_j \) in the graph in turn implies the absence of the factors \( P_i(X_i \mid \ldots X_j \ldots) \) or \( P_j(X_j \mid \ldots X_i \ldots) \) in \( P(\mathbf{X}) \), and thus the existence of a set of variables \( S \subseteq \mathbf{X} \setminus \{X_i, X_j\} \) that makes \( X_i \) and \( X_j \) conditionally independent in probability (Koller and Friedman 2009; Castillo et al. 1997).

The thereby obtained spatial structures provide a fingerprint that can be quantitatively analyzed by exploring the spatial distribution of the edges within a given GBN in terms of distance, or by the use of alternative similarity metrics described with detail in Graafland et al. (2020). The same article explains how the reliance of GBNs on equation (2) is determinant for the good quality of the fingerprint in the case of complex climate data, and where standard correlation networks, which would rely on equation (1), fall short.
In Section 3, this “edge distribution” will be analyzed for GBNs of four example GCMs and one reanalysis for illustrative purposes.

c. Probabilistic Querying: Evidence propagation in Bayesian Networks

In addition to the network structure’s capability to characterize local and remote spatial dependencies, the associated probabilistic model allows for probabilistic reasoning and querying (Castillo et al. 1997) by, e.g., introducing evidence at a particular location and then computing the resulting conditional probabilities for local and distant locations in the entire network, thereby quantifying the spatial (tele)connections associated with the point of evidence. Here, the JPD function of the BN can be used to estimate the impact of an evidential variable $X_e$ (at a given grid-box with known value) to other variables (at other grid-boxes). For example, assuming warming conditions in a particular grid-box of the globe $X_e$, e.g. a strong increase in temperature, say $X_e = 2\sigma_{X_e}$, the conditional probability of a given temperature anomaly at the other grid-boxes $P(X_i|X_e)$ is a quantification of the physical impact this evidence has on nearby or distant regions.

In practice this means computing conditional probabilities for a subset of variables given some evidence at a source variable triggering the dependencies. This problem is typically referred to as ‘evidence propagation’ or ‘inference’ in probabilistic graphical models. It differs from calculating conditional dependencies in a ‘general’ JPD function by the fact that we can benefit from the network structure of encoded conditional (in)dependencies allowing us to ‘propagate’ the evidence through the network, instead of the computational costly process of marginalizing out the evidence variables from the joint JPD. However, evidence propagation in BNs can also become computationally intensive (particularly for dense topological structures) and, in recent years, much investigation has been devoted to encounter practical solutions for different types of inference problems (Koller and Friedman 2009). In this study we use an approximate Monte Carlo approach in which the JPD function is estimated by random resamples of all or some of the variables in the network, so that they provide the best possible representation of the overall posterior (i.e., conditioned on the evidence) probability distribution. To this end, we use the likelihood weighting resampling method (Koller and Friedman 2009), designed to give more weight to samples closer to the posterior probability and thus more relevant to our evidence, and less weight otherwise.
Learning GBNs from data

Learning a GBN consists of two essential phases: In the structure learning phase the graph $\mathcal{G}$ is found that encodes the dependence structure present in the data. In the parameter learning phase the parameters $\Theta$ of $P$ are estimated. The graph structure of the BN identifies the parent set $\Pi_{X_i}$ in equation (2). With this structure available, one easily learns the corresponding parameter set $(\beta, \nu)$. In our case, the parameters $\beta_{ij}$ and $\nu_i$ are a maximum likelihood fit of the linear regression of $X_i$ on its parent set $\Pi_{X_i}$. To estimate the parameter values from the graph structure, we use the appropriate function in the R-package bnlearn (Scutari 2010). The challenge of learning the graph structure is explained in the following four paragraphs and parallel sections ‘learning BN structure (from data)’ in Graafland et al. (2020) with minor modifications.

The challenge of learning the graph structure is explained in the following four paragraphs and parallel sections ‘learning BN structure (from data)’ in Graafland et al. (2020) with minor modifications.

The graph of a BN is estimated with the help of a structure learning algorithm that finds the conditional dependencies between the variables and encodes this information in a DAG. Graphical (dis-)connection in the DAG implies conditional (in-)dependence in probability. From the structure of a BN, a factorization of the underlying JPD function $P(X)$ of the multivariate random variable $X$ (as given by Eq. (2)) can be deduced.

In general, there are three types of structure learning algorithms: constrained-based, score-based, and hybrid structure learning algorithms — the latter being a combination of the first two algorithms. Constrained-based algorithms use conditional independence tests of the form $\text{Test}(X_i, X_j | S; \mathcal{D})$ with increasingly large candidate separating sets $S_{X_i, X_j}$ to decide whether two variables $X_i$ and $X_j$ are conditionally independent. All constraint-based algorithms are based on the work of Verma and Pearl (1991) on causal graphical models, whose first practical implementation was seen in the Principal Components algorithm (Spirtes et al. 1993). In contrast, score-based algorithms apply general machine learning optimization techniques to learn the structure of a BN. To each candidate network a score reflecting its goodness of fit is assigned which the algorithm then attempts to maximize (Russell and Norvig 1995). In Scutari et al. (2019) we compared the three aforementioned algorithm classes in terms of accuracy and speed if applied to high-dimensional complex data and found the score-based algorithms to perform best, with the additional advantages of being able to 1) handle highly dimensional data with low sample size and 2) find networks of all desired sizes. Constrained-based algorithms, in turn, can only model complex data up to a certain size and, if applied to large climate datasets, only are able to reveal local network topologies.
Hybrid algorithms perform better than constrained-based algorithms on complex data, but worse than score-based algorithms.

In this work we use a simple score-based algorithm, the Hill Climbing (HC) algorithm proposed by Russell and Norvig (1995), to learn GBN structures. The HC algorithm starts with an empty graph and, in every iteration, tries to delete and reverse each arc in the current DAG. Moreover, it attempts to add each possible arc that is not already present in the current DAG and that does not introduce any cycles. Then the algorithm moves to the network with the highest score visited in this iteration, or the algorithm stops if no neighboring network with a higher score than the current network would have been found. In our case we used the Bayesian Information Criteria (BIC) score (referred to as BIC\(_0\) in Scutari et al. (2019)), which is defined as:

$$\text{BIC}(G; D) = \sum_{i=1}^{N} \left[ \log P(X_i \mid \Pi X_i) - \frac{|\Theta X_i|}{2} \log N \right],$$

where \(G\) refers to the graph (DAG) for which the BIC score is calculated, \(P\) refers to the probability density function that can be deduced from the graph (i.e. equation (2)), \(\Pi X_i\) refer to the parents of \(X_i\) in the graph (i.e. nodes \(Y\) with relation \(Y \rightarrow X_i\) in the graph) and \(|\Theta X_i|\) is the amount of parameters of the local density function \(P(X_i \mid \Pi X_i)\).

For all climatological datasets in this study the algorithm is stopped after 1800 iterations, in this way including around 1800 parameters/links. For this amount of parameters, Graafland et al. (2020) showed via cross-validation that the final DAGs are optimal in the sense of capturing both local and long-distance structures without redundancy, exhibiting a good balance between local and long distance dependencies that co-exist in the complex dataset from which the DAG is learnt. Note that the results obtained here do not substantially change for alternative assumed iteration numbers.

e. Probabilistic Distance Measures

The notion of distance in probability space allows us to calculate pairwise distances between JPDs. We can benefit from probability theory in the context of Bayesian networks to define for two example GBNs, GBN\(_P\) and GBN\(_Q\), each with their associated JPD function, \(P\) and \(Q\), a distance \(D(P, Q)\). Several measures of dissimilarity between pairs of Bayesian Networks BN\(_P\) and
BNQ have been tested in the present study. The first one is the commonly used Kullback-Leibler divergence
\[ D_{KL}(P||Q) = \int_{-\infty}^{\infty} p(x) \log(p(x)/q(x)) \, dx \] (Kullback 1959), typically used to determine
whether an observed distribution, Q, is a sample of another distribution, P. This definition, however,
requires the absolute continuity of the distribution function to be well defined. I.e. if \( p(x) \neq 0 \) then
\( q(x) \neq 0 \). The assumption is, however, generally not fulfilled by the Gaussian Bayesian Networks
applied in our study, which is why the Kullback-Leibler divergence was discarded.

Other distance measures that omit the the absolute continuity requirement are based on the
Bhattacharya coefficient, defined as \( BC(P, Q) = \int_R \sqrt{q(x)p(x)} \, dx \) (Hellinger 1909; Rust et al.
2010). This coefficient is symmetric with values between zero (P and Q have disjoint supports) and one (P and Q are identical). Roughly speaking, it can be thought of as a measure of overlap of
the two distributions. Dissimilarity, on the other hand, can be quantified by the negative logarithm
of the Bhattacharya coefficient, i.e. the Bhattacharya distance: \( d_B(P, Q) = -\log BC(P, Q) \) or by
the Hellinger distance, defined as \( d_H(P, Q) = \sqrt{1 - BC(P, Q)} \) (Hellinger 1909). At first glance,
the Hellinger distance seems favorable over the Bhattacharya distance, as it satisfies triangle
inequality— i.e. for JPDs P,Q and R we have \( d_H(P, Q) + d_H(Q, R) \leq d_H(P, R) \)— whereas the
Bhattacharya distance does not qualify as a true distance measure in this regard. However, in the
case of high dimensional distribution functions, the Bhattacharya coefficient is already a small
number and machine accuracy is not sufficient to distinguish the Hellinger distance from 1 for
all pairs of BNs. Therefore, the use of the Hellinger distance is not feasible in this work. Given
the above considerations, we finally decided to use the Bhattacharya distance to quantify the
dissimilarity between pairs of Bayesian Networks. Overall, it is the most comprehensive and
practical measure to work with in the context of the present study.

3. Results

a. Climate Probabilistic Networks: An Illustrative Example

In this section, we start with an illustrative example explaining the graph structure (Figure 1)
and the probabilistic model (Figure 2) of the GBNs learned from four illustrative CMIP5
GCMs and one reanalysis. The first row in Figure 1 shows the GBNs learnt from the four example
GCMs (BNU.ESM, ACCESS1.0, ACCESS1.3, and CMCC.CMS in column 1-4) and the reanalysis
(ERA-Interim, last column). In order to properly characterize the interplay of strong local and weak
long-distant dependencies between the grid points, each GBN is separated into two sub-networks in the second and third row, consisting of short and long-range interactions, defined by distances below and above 10000 km, respectively. For each GBN, the total amount of edges is shown in the first row and the sub-count for the two aforementioned distance classes is depicted in the second and third row, respectively.

As can be seen from the figure, BNU.ESM has many short-range edges (1747) and only a few long-range edges, hereafter also referred to as “teleconnections” (49). These numbers greatly differ from those obtained from the ACCESS models, comprising remarkably fewer short-range (∼1570/1580) and many more long-range edges (∼200). The spatial distribution of the edges in CMCC.CMS and ERA-Interim are similar to each other, comprising fewer long-range links than the ACCESS models but more than the BNU.ESM model. Note also that the teleconnections in ACCESS1.0 and 1.3 cover all parts of the world with a comparable number of zonal and meridional directions, whereas those in CMCC.CMS, BNU.ESM and ERA-Interim concentrate in the tropics.

Fig. 1. Visualization of the network structures for the BNs obtained from four illustrative GCMs and one reanalysis (ERA-Interim). The first row represents the whole network (numbers in the title indicate the number of links). The second and third rows represent the sub-networks of short (<10000 km) and long (≥10000 km) links, respectively, the latter characterizing teleconnection-like relationships.
Fig. 2. Composite maps of the differences between conditional and marginal probabilities for warm $P(X_i \geq 1 \mid X_e = 2) - P(X_i \geq 1)$ (red scale) and cold $P(X_i \leq 1 \mid X_e = 2) - P(X_i \leq 1)$ (blue scale) conditions modeled by the GBNs (the maximum of both quantities is displayed in each gridbox with the associated colorbar). The location of the evidence variable $X_e$ is signalized with a white box in the different panels. The event $X_e = 2$ indicates a positive anomaly of the monthly mean temperature in excess of two standard deviations, indicating strongly anomalous warm conditions for $X_e$. The evidence in the first row represents a warm anomaly in a gridbox in the central Pacific (emulating El Niño conditions), whereas the evidence given in the second row represents a warm anomaly in the southern Pacific.

and are dominated by zonal directions, which is consistent to the well-known general temperature increase in the entire tropical belt during El Niño events (Trenberth et al. 1998; Brands 2017).

Besides the network structure, a GBN learns a probabilistic model associated with the dependency structure encoded in the network. This allows for probabilistic querying (see Section 2c) the model to obtain relevant statistical information. This is illustrated in Figure 2 visualizing the probability temperature anomaly pattern conditioned on a given evidence of positive temperature anomaly at a certain grid-box $X_e$. The figure shows the results for two illustrative evidence gridboxes, one in the center of the Niño 3.4 region (marked with a white box in the maps in the first row), and another in the extra-tropical South Pacific (marked with a white box in the maps in the second row). Both regions are known to be related with the El Niño-Southern Oscillation (ENSO) teleconnection patterns.

Given the anomalously warm conditions used as evidence, the conditional probabilities of a warm or cold anomaly magnitude equal or greater than one standard deviation are calculated at all other grid-boxes (i.e., $P(X_i \geq \sigma_{X_i} \mid X_e = 2\sigma_{X_e})$ and $P(X_i \leq \sigma_{X_i} \mid X_e = 2\sigma_{X_e})$, respectively). This is done
separately for each of the four example GCMs and the single reanalysis. In the reference reanalysis (first row, fifth column in Figure 2), the well-known ENSO teleconnection patterns documented in many previous studies are here detected as well (Halpert and Ropelewski 1992; Hoerling et al. 1997; Trenberth et al. 1998; Wallace et al. 1998; Brands 2017; Domeisen et al. 2019). These include the typical tripole pattern of elevated probabilities for positive anomalies located in the central-to-eastern equatorial Pacific and along the western coastlines of sub-tropical North and South America, surrounded by a boomerang-like pattern of elevated probabilities for cold anomalies in the western-to-central sub-tropical Pacific of both hemispheres. Elevated probabilities for positive temperature anomalies are also obtained in eastern Australia and in the North and South Pacific at mid-latitudes, roughly coinciding with the respective branches of the Pacific North American and Pacific South American patterns in these regions, i.e. with the Aleutian and Amundsen Sea low pressure systems (Wallace and Gutzler 1981; Barnston and Livezey 1987; Mo and Ghil 1987). At long-range, elevated probabilities for positive temperature anomalies are detected in the equatorial Indian Ocean and surrounding land areas, particularly the Mozambique region, and also in the sub-tropical South Atlantic off-the-coast of Brazil.

The quasi observed conditional probabilities in the tropical and sub-tropical Pacific Ocean (in the first row of Figure 2) are overestimated by BNU.ESM and CMCC.CMS and underestimated by ACCESS1.3, with ACCESS1.0 being closest to ERA-Interim from visual inspection. A similar picture is obtained for the Indian Ocean. The much weaker teleconnections with the extra-tropics seen in observations are generally best reproduced by CMCC.CMS and poorest results for this case are obtained with ACCESS1.3.

The findings related to the second case of evidence propagation, i.e. prescribing anomalously warm conditions in the subtropical central South Pacific Ocean, are presented in the second row of Figure 2 and support the aforementioned results. The reference reanalysis (second row, fifth column) reveals teleconnection patterns akin to those found in the initial case. However, the positive anomaly probabilities in the equatorial Indian Ocean are lower than in the first case. Notably, a circlelike quadrupole pattern emerges in the reanalysis comprising negative anomaly probabilities around New Zealand and off the Chilean coast separated by positive anomaly probabilities in the subtropical to mid-latitude central South Pacific and equatorial Pacific. This pattern is most accurately replicated by CMCC.CMS. BNU.ESM, however, generally flattens out the quadrupole’s
magnitude, with too weak cold anomaly probabilities and too ample warm anomaly probabilities in the tropical Pacific extending too far to the West. The ACCESS models capture the subtropical part of the quadrupole (i.e. the tripole of cold-warm-cold anomaly probabilities located there) but underestimate the extension of warm anomaly probabilities in the equatorial Pacific, which are completely missing in ACCESS1.3. The ACCESS models also simulate unrealistically large probabilities for cold anomalies in the southern Indian Ocean if compared to the reference reanalysis.

At this point, it is worth noting that the teleconnections related to ENSO focused on here so far are only one phenomenon that has been successfully learnt by the GBNs. Indeed, however, the probabilistic network has learnt the simultaneous teleconnections seen in surface air temperatures triggered at every region of the world and thus contains far more information than those shown above.

b. Quantifying Model Similarity: An Illustrative Example

In this section we illustrate different approaches to quantify the similarity between two probabilistic networks using the illustrative example introduced in the previous section. We show the use of both network- and probabilistic-based metrics for this purpose. In particular we illustrate the Bhattacharyya distance which is used throughout the rest of the paper.

A simple similarity method can be defined based on the network topologies. In particular, Figure 3(a) shows the average minimum amount of links needed by a candidate GBN (Network 1, rows) to cover a long-range link between two nodes of another GBN (Network 2, columns)\(^1\). The numerical results of Figure 3(a) agree with the qualitative results obtained from Figure 1. Namely, we see that ERA-Interim and CMCC.CMS perform similarly in terms of the long-distance-coverage-measure. Both need relatively few edges to cover the long-range links of BNU.ESM, but many edges to cover the complex global dependency structure of the ACCESS models. Due to the low number of own long-distance links, BNU.ESM on average needs many short range edges organized in v-structures to cover the more frequent long-range links in the four remaining datasets. Conversely, the ACCESS models, containing the largest number of long-range edges in this example, need few links to cover the far less frequent long-range edges present in BNU.ESM. Interestingly, however, they do not

\(^1\)This measure is here applied on the moral graphs of the GBNs in Figure 1 in order to include links between variables that are marginally dependent but, due to their conditional dependence structure, are connected by a v-structure passing through an intermediate third variable.
Fig. 3. Illustration of different approaches to quantify the similarity between probabilistic networks using (a) network- and (b) probabilistic-based metrics. Panel a shows the results for the long-range links coverage from the BNs of the example subset. Each entry of the matrix presents the average amount of links that is needed in network 1 to cover a random link of more than 10000 km in network 2. Panel b shows the Bhattacharya distances between the BNs of the example subset. Each entry displays the symmetric Bhattacharya distance between BN 1 and BN2. Small distance values indicate similar spatial dependency patterns; high distance values indicate the opposite.

Differ much from CMCC.CMS in reproducing the long-range edges from ERA-Interim although CMCC.CMS clearly comprises less long-range links, meaning that the direction/orientation of the links in the latter GCM are closest to those obtained from quasi-observations (ERA-Interim). CMCC.CMS is thus the best performing GCM in this example. We also conclude that the inclusion of many long-range edges does not guarantee the correct representation of any reference long-range spatial pattern if the edges’ localization does not match, leading to wrong orientations in the visualized graphs. Network topology is thus not only determined by the the amount but also by the localization of the long-range edges. An overview of other indices that can be derived from network topology is provided in Graafland et al. (2020).

A more comprehensive approach can be defined building not only on the network structure but also on the underlying probabilistic model (illustrated in Figure 2), thus fully accounting for the information contained in the GBNs. In this case, we use the notion of distance in probability space, allowing us to calculate pairwise distances between JPDs. To this end, we use the Bhattacharya
distance as a probabilistic distance measure between two JPD functions (Kailath 1967), each associated with a specific GBN as defined in Section 2e. Figure 3(b) shows the Bhattacharya distance (BD) between the example subset treated in the upper sections. The aforementioned probabilistic and topological similarities are captured well by this symmetric metric. As was qualitatively described above, the ACCESS models building on similar atmospheric sub-models comprise very similar spatial edge distributions, which is reflected by a small BD value. Their topology structures clearly differ from that obtained for CMCC.CMS which, in turn, is similar to ERA-Interim, translating into large or small BD values respectively. BNU-ESM is the most distant model of this example subset and consequently receives large BD values for any pair-wise comparison, particularly when compared with the two ACCESS models.

c. Analyzing model similarity in the CMIP5 ensemble

In the previous sections, we illustrate how GBNs encode the spatial dependency structures learnt from GCM model outputs and show that differences among the resulting GBNs are effectively described by the Bhattacharya Distance (BD). A small BD corresponds to similar spatial dependency structures in the data which is used here as an indication of potential similar model formulation (model similarity). We calculate in this section the Bhattacharya distance between all possible pairs of the full multi-model ensemble (29 GCMs) and the two reanalysis considered in this study. Results are displayed in Figure 4 and model dependency is here defined in a posteriori context: Rather than complete dependency or independency designation we consider a continuous range from highly dependent to very independent models —e.g. from very low Bhattacharya distance to very high Bhattacharya distance— guided by the eight levels (eight quantiles) of BD as indicated in the colorbar.

The GCMs in Figure 4 are ordered as follows according to their BD distances. First, the Euclidean distance between rows of the Bhattacharya distance matrix is calculated. Then, a hierarchical clustering method is applied that initially assigns each row (except the first two representing reanalyses) to its own cluster and then iteratively joins the two most similar clusters following the complete linkage method. I.e. the two clusters that have the shortest distance between their furthest elements are joined in each iteration. Finally, a single cluster remains. Similar results have been obtained with alternative clustering methods such as “k-medoids” (Kaufman and Rousseeuw
The GCMs are ordered according to the results obtained from a hierarchical clustering of the results in the matrix. The associated dendogram is displayed at the top and cut off at the red dashed line. Groups of clustered models below the cut off level are assigned a colored box. Red blocks indicate GBNs whose GCMs are produced by the same institute. Red blocks with a cross are grouped above the cut off level and indicate BNs built upon GCMs from the same institute, but with substantially differing AGCMs, as documented in Knutti et al. (2013); Boé (2018); Brands et al. (2023)). Purple blocks indicate GBNs whose GCMs share a significant amount of their atmospheric model component. The orange dashed boxes represent GBNs with undocumented similarities in their GCMs.

![Diagram showing Bhattacharya distances for all possible combinations between the 29 GCMs and two reanalyses. The GCMs are ordered according to the results obtained from a hierarchical clustering of the results in the matrix.](image-url)
In this way, those GCMs with similar distances to all other models are located close to each other in the distance matrix. The dendogram on the top illustrates the hierarchical clustering process and the red dotted line represents the cut off below which clustered groups are considered interdependent models.

As was the case for the example ensemble in Section 3b, the diagonal of the matrix is zero, as expected, since the distance between equal GBNs is zero. Moreover, models of the same institutes are generally clustered and are easily recognized by remarkably low pairwise distance values if the AGCMs used therein do not differ from each other (e.g. red boxes). Some examples are MIROC.ESM.CHEM and MIROC.ESM or HadGEM.CC and HadGEM2.ES, with values of 18 and 17, respectively. Relatively large distances between models from the same institutes are associated with substantial changes in the atmospheric sub-components of the models. They are assigned a red box with a cross in Figure 4. Note, for example, the close similarity between the pair GFD-ES2M2 and GFDL-ESM2M on the one hand and their relative distance to GFDL-CM3, comprising differences in the atmospheric sub-model as described in Donner et al. (2011). Another example is the low distance between the pair IPSL.CM5A-LR and IPSL.CM5A-MR comprising similar parametrization schemes in the atmosphere on the one hand, and the single model IPSL-CM5B-LR using substantially modified schemes if compared to the former two on the other (Dufresne et al. 2013).

Likewise, small Bhattacharya distances ($BD = 30$) are found between models of different institutions which use the same atmospheric sub-model or versions thereof (assigned purple boxes), such as CNRM.CM5 and EC.EARTH relying on ARPEGE/IFS and ECWMF/IFS respectively, jointly developed by Meteo-France and ECMWF (Voldoire et al. 2013)). Similarly low distances are obtained for CMCC-CMS and MPI-ESM-LR ($BD = 26$) both relying on ECHAM versions in the atmosphere, and for ACCESS1.0, ACCESS1.3, HadGEM2-ES and HadGEM2-CC based on versions of the HadGAM2 AGCM ($BD \in \{30,31\}$), as well as NorESM, CESM1.BGC and CCSM4 relying on CAM versions in the atmosphere. These results are similar to those obtained in Brands (2022b) and Merrifield et al. (2023) for completely different phenomena and distance measures (i.e. error analysis of Lamb weather types for a delimited region versus global climatological surface temperature and sea pressure level fields, respectively), showing that the AGCM is the most important determinant of model similarities in the atmosphere.
Notably, some previously undocumented GCM similarities are detected in the present study. Most apparent is the big cluster comprising 15 GCMs (from IPSL-CM5A-LR to GFDL-ESM2M) with relatively low Bhattacharya distances containing all GCMs developed in North America (except GFDL-CM3), namely NorESM1-M, CCSM4, CESM1-BGC, CanESM2, GFDL-ESM2M, GFDL-ESM2M. Moreover, HadGEM2-CC/ES and ACCESS-1.0/1.3 are relatively close to CSIRO-MK3.6, probably due to their shared modeling history (Bi et al. 2013). Surprisingly, CSIRO-MK3.6 is also close to CNRM-CM5 and CMCC-CMS, although the former does not share a single sub-model with the latter two (Brands et al. 2023). EC-Earth and MIROC as well as MRI.CGCM3 and CMCC.CMS are also remarkably similar despite the distinct component models in use, pointing to convergence of conceptually distinct models as outlined in Boé (2018). The above mentioned and other undocumented but close GCMs that are clustered beneath the red dotted line are marked with an orange box in Figure 4.

Finally, the distances between the two reanalyses is indicated by the dark blue box. A very low distance (26) is obtained for the reanalysis pair (ERA-Interim and JRA-55), which can be interpreted mainly as observational uncertainty for the probabilistic networks obtained here (Brands et al. 2012). This kind of uncertainty is smaller than the distance (or error) of any GCMs w.r.t to ERA-Interim or JRA-55—one would expect to obtain a few GCM vs. reanalysis distances to be smaller just by chance—, meaning that a change in the underlying reference reanalysis does not substantially change the results.

From a methodological perspective, there could be many reasons for the unexplained GCM similarities found above, but they should have some common constraint (as with reanalysis and observations) that drives the similar spatial structure detected with GBNs; whether this should be considered model dependence or not is still an open question.

Part of the information captured by the Bhattacharya distance in the matrix is illustrated in Figure 5, showing the conditional probabilities of receiving a temperature anomaly $\pm$ one standard deviation all around the globe given a positive anomaly of two standard deviations at the “triggering” grid-box in the central tropical Pacific. The maps are grouped according to the dendogram learnt from the Bhattacharya distances in Figure 4 and the distinct colors point to the distinct groups obtained from this technique. Albeit the Bhattacharya distance used here measures similarities between the full (or global) dependency structures of the considered GCMs or reanalyses, the
Fig. 5. Differences in the conditional and marginal probabilities (as in Figure 2) modeled by the GBNs learnt from two reanalyses and 29 GCMs. The location of the evidence is the same as in the first row of Figure 2. The maps are grouped according to the dendogram displayed in Figure 4 and are assigned the same colour frames: Red blocks indicate GBNs whose GCMs are produced by the same institute. Purple blocks indicate GBNs from GCMs that share a significant amount of their atmospheric model component. The orange dashed boxes indicate GBNs from GCMs grouped by the clustering method for other reasons. Finally, maps with stars correspond to the single leaves.
obtained values are representative of the similarities in the spatial patterns of the conditional probabilities triggered by ENSO only, showing that these are most important within the learnt dependency structures. Finally, the Bhattacharya distance also correctly identifies the few outlier models present in the ensemble (inmcm4 and BCC-CSM1.1).

d. Effect of internal variability

In this section, we analyze the influence that internal variability might have on the spatial dependency pattern of a single GCM. The internal variability is inherent to the stochastic nature of the climate system and can be empirically characterized using a set of simulations from different initial conditions. For instance, the CSIRO.Mk.3.6.0 model provides an ensemble of ten distinct initial condition simulations as part of its contribution to CMIP5. We analyze the probabilistic networks resulting from this ensemble following the same GBN methodology introduced in the previous sections.

The upper block of Figure 6 shows the Bhattacharya distances between the ten available runs of the CSIRO model in CMIP5. We observe that differences in spatial patterns due to internal variability in the CSIRO ensemble are small, reflected by a Bhattacharya distance of BD = 17 with slight variations among members. The lower block of Figure 6 shows the distance of the initial condition runs with respect to the other CMIP5 models. All CSIRO models have a similar distance to other CMIP5 models, with the next-closest model to the CSIRO runs being CNRM.CM5 (BDs of ~ 34), followed by the ERA-Interim reanalysis (BDs of ~ 37).

Figure 7 shows the effect of internal variability in the resulting conditional probability patterns, given El Niño related evidence, as in Figure 2. The resulting probability patterns exhibit a negligible impact of internal variability. This figure highlights another important aspect, since the different sequences of El Niño events generated in the different runs are not determinant for the robustness of the method, nor the number of any other determinant climate modes (as reflected by the BD distances, in which all climate modes are captured). This shows that the period of 30 years is sufficiently large to determine characteristic GCM fingerprints with GBNs.
Fig. 6. Bhattacharya distances for all possible combinations between (columns) the ten initial condition runs of CSIRO and (rows) the 28 CMIP models (alphabetically ordered) and two reanalyses.

4. Discussion and Conclusions

In this work, the global spatial dependency structure of monthly near-surface temperature is analyzed for the historical experiments of 29 GCMs used in CMIP5 and for two reanalyses,
Fig. 7. Differences in the conditional and marginal probabilities (as in Figure 2) modeled by the GBNs learnt from the ten initial condition runs of CSIRO.MK3.6.0. The location of the evidence variable $X_e$ is the same as in the first row of Figure 2. The event $X_e = 2$ indicates a positive temperature anomaly of two standard deviations above the mean value.

considering the period 1981-2010. To this end, Gaussian Bayesian Networks (GBNs) are applied, which are capable to extract the manifold spatial dependencies present in these data in a purely data-driven manner, i.e. by means of graphical links and parameters of the Joint Probability Density function that vary in strength, distance and location. The learnt networks in principle contain independent information about all spatial dependencies present in the data, which is illustrated for an example subset of four GCMs and a single reanalysis.

In each network a large variety of spatial dependencies is found which generally become sparser and weaker with increasing distance between parents and children. Remarkably, the structures learnt from some GCMs largely differ from those learnt from the reanalysis data, particularly because of many more long-range dependency structures present in the former. In the example subset, the method’s meaningfulness is further illustrated by explicitly calculating the conditional probabilities associated with ENSO events, obtaining teleconnection patterns that are in close agreement with those documented in previous studies relying on classical methods such as correlation and composite analysis (Halpert and Ropelewski 1992; Domeisen et al. 2019).

The Bhattacharya distance has then been applied to measure inter-model differences in the dependency structure and its associated probabilities in an objective manner. It measures structural similarities and differences between a pair of GCM realizations. Less distance indicates a similar
strength and spatial structure of the variability produced by the two candidate GCMs. In essence, the Bhattacharya distance measures to which degree these realizations have been drawn from the same underlying chaotic system and it is here applied to all possible GCM and reanalysis combinations.

By applying a hierarchical clustering algorithm on the obtained GCM distance matrix, differences are shown to be associated with distinct architectures of the atmospheric model components in use. Other factors, such as the architectures of the remaining model components (ocean, land-surface etc.), or distinct responses to the common external forcing, also likely play a role for creating distances. The simulation of important climate modes, such as the El Niño Southern Oscillation, can be used to explore the differences between GCMs, and can help identifying dissimilarity factors. Finally, the Bhattacharya distance can be used to weigh the distinct GCMs according to their similarities.

Many of the here obtained results are in line with those obtained in earlier studies on model similarities. The classification of CMIP3 and 5 GCMs in Boé (2018) was based on climatological averages. In that work, a clear relationship between the number of shared model components and the proximity of their realizations was demonstrated. A drawback of this analysis is that it did not take into account (spatial) variability, which is a crucial characteristic of the climate system’s chaotic nature. Model development often focuses as much on variability as on the mean state (Covey et al. 2000).

Brands (2022b) classified 61 GCMs from CMIP5 and 6 according to the spatial pattern of their error in reproducing regional-scale atmospheric circulation patterns in the Northern Hemisphere extra-tropics as described by reanalysis data. As in the present study, GCMs were found to produce similar spatial error patterns if they use similar atmospheric sub-models and the CMIP ensemble was found to be surprisingly dependent overall, obtaining an average spatial error correlation coefficient of +0.6 on average, when taking into account all possible GCM comparisons. Albeit distinct phenomena are assessed, the spatial correlation analysis conducted in Brands (2022b) seems to be less discriminant than the approach followed in the present study, the latter returning clearer differences in the pair-wise GCM dependencies.

Masson and Knutti (2011b) and Knutti et al. (2013) assessed GCM dependencies by fitting a multivariate Gaussian distribution to spatial temperature fields, using the Kullback-Leibler diver-
gence to measure dissimilarities between them. To avoid singularity problems otherwise caused by the inverse sample covariance matrix, they reduce the spatial dimensionality by projecting the GCM and reanalysis data on a new coordinate system based on the empirical orthogonal functions obtained from the reanalysis datasets. This change of the coordinate system, however, is somewhat random for the GCM datasets and implies the loss of the geographic reference system and thus the loss of spatial interpretability. The method allows for quantification of distances between spatial fields but gives no indication on why and where these distances occur.

Brunner and Sippel (2023) go further including non-linear spatial patterns in their analysis with CNNs, removing the seasonal mean and global mean from daily temperature datasets. They show that out-of-sample realisations can accurately be assigned to the correct GCM model in 83 per cent of the cases. Misclassifications (assigning a realisation of model A to model B) can be interpreted as similarity between those models and occur mostly between models developed in the same institution or between models for which shared model components are documented. One of the current challenges of deep learning methods is the lack of easy interpretability of its results. CNNs do not allow a straightforward extraction of the features used to separate categories. This has given rise to the currently active field of investigation hence of eXplainable Artificial Intelligence (XAI) in atmospheric sciences (Barnes et al. 2020; McGovern and Lagerquist 2020; González-Abad et al. 2023; Silva and Keller 2023). E.g. the work of Bach et al. (2015), proposing a general solution to the problem of understanding classification decisions by pixel-wise decomposition of nonlinear classifiers, is promising in the context of classifying spatial dependency structures (Brunner and Sippel 2023). With respect to deep learning techniques, GBNs have the advantage that they natural provide a ‘pixel-wise decomposition’ building on the closed-form solution of the JPD as a factorization of the variables in conditional probability density functions.

The classification method applied in Nowack et al. (2020) is based on constructing causal networks that are consecutively compared on a link-to-link basis. Varimax PCA was employed to reduce the spatial dimensions of the datasets, in this case, to select the most significant regions in the reanalysis dataset used as variables in the causal networks. This type of dimension reduction retains the original coordinate system and, unlike in Boé (2018); Knutti et al. (2013); Masson and Knutti (2011b), and also in Brunner and Sippel (2023), has the advantage that the derived causal networks can be qualitatively interpreted. Nowack et al. (2020) also mentions that causal networks
can potentially be applied to explore Earth system dynamics, such as global teleconnections and their associated directions. A drawback to their method, however, is still the need for dimension reduction. To apply causal networks to complex systems data like climatological datasets, a reduction of dimensionality is necessary. The (variants of) causal structure learning algorithms applied in Nowack et al. (2020) are applied up to today to datasets that include orders of ten variables. More on the ‘curse of dimensionality’ in causal network analysis is explained in Runge et al. (2019) and Runge et al. (2023). This limited number of variables compels a choice between focusing solely on global-scale patterns (via dimension reduction, often partly expert-driven instead of fully data-driven) or exclusively on regional granular analysis (omitting parts of the globe and thereby reducing dimension).

Generally speaking, Gaussian Bayesian Networks have the great advantage of being able to combine both of the two main approaches followed in aforementioned studies: the statistical and the network approach. This twofold approach optimizes the amount of quantitative and qualitative information that can be extracted from climatological datasets (reanalyses and GCM runs in this case) and provides insight to the causes of pair-wise similarities or differences. On the one hand, network analysis gives insight into the amount of variability present in the GCM integrations. On the other hand, probabilistic analysis indicates to which degree the spatial structures of this variability differ from one model to another and whether there exist overlapping dependency structures. Finally, the fine granularity on global scale –with respect to the above studies– enables the observation of both local and global processes in the data, facilitating an analysis of the climate system as a complex system in which emergent phenomena occur. This approach is advantageous in the context of model dependency, as model components in GCMs capture both global and local processes and their interplay.

Albeit the hierarchical clustering algorithm applied to the pair-wise Bhattacharya distances tends to put GCMs using similar atmospheric sub-models into the same group, which is in line with all above mentioned studies, there are exceptions from this rule in which nominally different AGCMs are grouped together, as mentioned before. This points to the relevance of other climate system components for the atmosphere, one promising candidate being the ocean, particularly at low latitudes, and to the possibility of GBNs to detect complex dependency structures mixing short-
range and long-range connections that are difficult to extract with traditional methods or less easy to interpret with advanced machine learning methods.
Acknowledgments. We acknowledge the three anonymous referees for their constructive comments which helped to improve this manuscript. This study is part of the R&D project “Eventos extremos compuestos para la evaluación de los impactos del cambio climático en la agricultura” (COMPOUND: TED2021-131334A-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. SB was funded by the European Commission – NextGenerationEU (Regulation EU 2020/2094), through CSIC’s Interdisciplinary Thematic Platform Clima (PTI Clima)/ Development of Operational Climate Services. JMG would like to acknowledge the ATLAS project (PID2019-111481RB-I00) funded by MCIN/AEI (doi:10.13039/501100011033)

Data availability statement. The monthly-mean near-surface air temperature data from the two reanalysis and 29 GCMs participating in CMIP5 are publicly available through ESGF (https://esgf-data.dkrz.de/) under the Creative Commons Attribution license CC-BY 4.0. User-friendly access to ESGF datasets and a variety of remote climate data sources is provided through the User Data Gateway, an integration of climate4R (Iturbide et al. 2019) with the Santander Climate Data Service (SCDS) THREDDS Data Server. This service is maintained by the Santander Meteorology Group (University of Cantabria - CSIC). The maps in figures have been created using the R-package visualizeR v1.5.1 that forms part of climate4R.

References


35


Accepted for publication in Artificial Intelligence for the Earth Systems. DOI 10.1175/AIES-D-23-0073.1.


