Bayesian structure learning for climate model evaluation

Terence John O'Kane[1], Dylan Harries[2], and Mark A. Collier[1]

[1]Commonwealth Scientific and Industrial Research Organisation (CSIRO)
[2]Commonwealth Scientific and Industrial Research Organisation

January 19, 2024

Bayesian structure learning for climate model evaluation

Terence J. O'Kane[1], Dylan Harries[2], Mark A. Collier[3]

[1]CSIRO Environment, Battery Point, Australia
[2]South Australian Health and Medical Research Institute (SAHMRI), Adelaide, Australia
[3]CSIRO Environment, Aspendale, Australia

Key Points:

- Bayesian structure learning is used to quantify uncertainty in estimated network structures describing climate mode teleconnections
- Dynamic Bayesian networks estimated from reanalyses are compared to CMIP5 model simulations over the historical period
- Differences in network structures between models and reanalyses quantify complex interacting biases in climate model dynamics

Corresponding author: Terence O'Kane, terence.okane@csiro.au

Abstract

A Bayesian structure learning approach is employed to compare and contrast interactions between the major climate teleconnections over the recent past as revealed in reanalyses and climate model simulations from leading Meteorological Centers. In a previous study, the authors demonstrated a general framework using homogeneous Dynamic Bayesian Network (DBN) models constructed from reanalyzed time series of empirical climate indices to compare probabilistic graphical models. Reversible jump Markov Chain Monte Carlo (RJMCMC) is used to provide uncertainty quantification for selecting the respective network structures. The incorporation of confidence measures in structural features provided by the Bayesian approach is key to yielding informative measures of the differences between products if network-based approaches are to be used for model evaluation, particularly as point estimates alone may understate the relevant uncertainties. Here we compare models fitted from the NCEP/NCAR and JRA-55 reanalyses and CMIP5 historical simulations in terms of associations for which there is high posterior confidence. Examination of differences in the posterior probabilities assigned to edges of the directed acyclic graph (DAG) provides a quantitative summary of departures in the CMIP5 models from reanalyses. In general terms the climate model simulations are in better agreement with reanalyses where tropical processes dominate, and autocorrelation time scales are long. Seasonal effects are shown to be important when examining tropical-extratropical interactions with the greatest discrepancies and largest uncertainties present for the Southern Hemisphere teleconnections.

Plain Language Summary

Climate model biases and performance is typically assessed against observational products via systematic comparison of individual metrics, usually focused on the mean climate, over the recent historical period. We demonstrate how Bayesian structure learning can enable a systematic probabilistic framework for process-based model evaluation of both the temporal behaviour of individual climate modes but also to identify and assess the teleconnections between those modes. We show that network structures can be fitted simultaneously and feasibly across a representative sample of climate model simulations affording uncertainty estimation of the robustness of differences across models and observations and robustly identify model biases between teleconnections in the climate.

1 Introduction

Bayesian methods allow for explicit estimation of uncertainties making them a natural choice for data analysis in situations where there are multiple sources of uncertainty and limited data. In this study we are motivated to implement Bayesian inference for climate model evaluation in terms of networks, often referred to as structural causal models, to understand how biases interact, as compared to observational networks that are themselves uncertain.

Climate model evaluation is typically conducted in terms of any given model's ability to accurately reproduce the observed climatological values and variations of the diverse processes that define the Earth's climate. For example, a model's climatology for individual fields such as sea surface temperature (SST) or mean sea level pressure (MSLP) may be compared to observations to characterize biases in the overall time-mean state. On smaller scales, the fidelity with which models reproduce particular, localized modes of variability may provide some indication of the reliability of projections relating to these modes. Deficiencies in the modelling of any individual mode may in turn propagate, via (causal) physical interactions, to manifest as biases in the representation of other teleconnections. While the detailed mechanism will be dependent on the complex dynamics present in the coupled system, recent approaches that represent this system in terms of a network of a relatively small number of interacting modes (Tsonis & Roebber, 2004; Tsonis & Swanson, 2008; Tsonis et al., 2008; Donges et al., 2009b, 2009a; Steinhaeuser et al., 2011, 2012) can provide an intuitive, albeit highly simplified, description of the key physical processes. In particular, learning the structure of the interactions between teleconnections in a model and comparing the results to similar structures inferred from observations provides a means of assessing the model's representation of coupled modes of variability (Falasca et al., 2019; Nowack et al., 2020; Vázquez-Patiño et al., 2020).

To ascertain the utility of any given climate model projection in this way requires an observational estimate against which model biases may be quantified. In climate science, reanalyses or state estimates (Kalnay et al., 1996; Onogi et al., 2007; Kobayashi et al., 2015; O'Kane et al., 2021) are typically used as proxies for the true history of the climate over the recent past. To produce a reanalysis, a climate model is constrained using available observations via formal data assimilation methods to estimate the true trajectory of the state of the climate, including the observed relationships between the ma-

jor modes of variability. However, the quality of any given state estimate will be impacted by factors such as the biases inherent in the particular climate model, the type of assimilation scheme, and the quality and spatio-temporal distribution of the available observations, including random variability. Consequently, the reanalysis datasets against which free-running models are evaluated will themselves almost certainly contain significant uncertainties.

Comparisons of different reanalysis products using any of the proposed network-based approaches highlight some of these issues. For example, the dynamics of major climate drivers within a given reanalysis may be summarized in a highly simplified fashion by fitting linear vector autoregressive models to timeseries of empirical indices characterizing the modes of interest. This model can be graphically represented using a directed acyclic graph (DAG), with nodes corresponding to the time-lagged indices and edges to the inferred Granger causal relationships (informally, a given variable is Granger causal to another if better predictions of the second are obtained by the inclusion of information about the first in comparison to predictions where this information is withheld). For climate data this makes readily achievable quantitative tests for identifying such relationships from observational data that would be superior to simpler lagged regression approaches (McGraw & Barnes, 2018; Bach et al., 2019). When the probability density function (PDF) for each index, conditional on the prior values of all other indices, is specified, the resulting model can be regarded as a dynamic Bayesian network (DBN). Applying this approach to two reanalysis products, namely JRA-55 (Kobayashi et al., 2015) and NNR1 (Kalnay et al., 1996), demonstrates good qualitative agreement overall (Harries & O'Kane, 2021). Systematic differences between the two products tend to coincide with known biases in the climate models underpinning each analysis. In some cases, however, differences between the single best-fitting model for each of the two products are found to involve edges that, given the data, would have low probability when considering the space of possible models. Estimation of these probabilities through the use of Bayesian methods allows uncertainties in the inferred relationships to be quantified. Attention can then be focused on those differences for which there is robust evidence (for example, the presence of an edge in the graph with high posterior probability in one product and its absence with equally high confidence in the other), which may suggest genuine biases.

Utilizing the above approach for model evaluation requires comparing differences between modeled and observed networks in much the same way, including evaluation of the robustness of any differences found. Within the context of an observational record that limits the temporal extent of reliable CMIP5 historical model and reanalysis data, Bayesian methods allow uncertainties in the estimated graphical models to be quantified conditional on the available observed data and modeling assumptions. Posterior probabilities for the presence or absence of a particular edge in the network may be estimated by sampling from the posterior distribution over allowable networks. When a relationship between modes is present in the model being evaluated with high confidence and absent in reanalysis data (or vice versa), we might expect this difference will be robust and reflect model bias. That is, even allowing for uncertainty in the best fitting network structures in the model and reanalysis, it is likely that this difference is present. Marginal posterior probabilities for relationships of interest in models and observations thus provide an intuitive and physically interpretable means of identifying model biases. For small networks, or at significant computational expense, calculation of these probabilities can be done by employing score-based structure learning methods for graphical models (Heckerman et al., 1995; Arnold et al., 2007; Lèbre, 2009). With a choice of conjugate priors, closed-form expressions are available for the required posterior densities, permitting efficient sampling (Geiger & Heckerman, 1994).

Here we apply these methods to evaluate the relationships between several global climate drivers obtained from CMIP5 models from seven of the leading meteorological centers in relation to the aforementioned JRA-55 and NNR1 reanalyses. The models (HadGEM2-CC (Martin et al., 2011); CanESM2 (Yang & Saenko, 2012); CNRM-CM5 (Voldoire et al., 2013); MIROC5 (Watanabe et al., 2013); ACCESS1-0 (Bi et al., 2013); NorESM1-M (Bentsen et al., 2013); GFDL-ESM2M (Dunne et al., 2013)) were chosen on the basis of having been assessed against performance criteria and selected from the 40 available CMIP5 models for inclusion in the "Climate change in Australia program" [1] for providing the requisite data required to calculate the major teleconnection indices; and to provide a representative sample of the current state of the art in climate models. Our explicit aim is to more generally demonstrate the utility of using Bayesian statistical methods for identifying robust relationships and uncertainty quantification between climate

[1] https://www.climatechangeinaustralia.gov.au/en/

teleconnections from data in a systematic manner as an additional approach to inform climate model development.

In section 2 we describe the CMIP5 models and reanalysis data to be interrogated and the methods used to reduce that data to a set of diagnostic timeseries describing the major atmospheric teleconnections and intraseasonal to interannual modes of climate variability. Section 3 describes the formulation of the Bayesian network model and choice of priors followed in section 4 by the resulting DAGs and quantitative comparisons of posterior distributions across models. A summary of our conclusions are in section 5.

2 Data and diagnostics

The reanalysis data that we analyze are obtained from the Japanese 55-year Reanalysis (JRA-55) and the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) Reanalysis 1 (NNR1) (Kalnay et al., 1996).

The NCEP/NCAR Reanalysis 1 is an atmospheric reanalysis covering the years 1948 to present. The data assimilation system employs a global spectral model with a T62 resolution on 28 vertical levels, and assimilates surface and atmospheric observational data. While a fixed analysis and forecast system is used for the duration of the reanalysis, changes in observing systems, notably a steep increase in satellite observations through the 1970s, still have an impact and, consequently, the reanalysis is less reliable in the first decade than at later times (Kistler et al., 2001). NNR1 represents a first generation reanalysis providing a multidecadal record of the atmospheric state, albeit with several known errors (Kistler et al., 2001) and biases, particularly in data-sparse regions in the high latitudes and the Southern Hemisphere (SH) (see, e.g., Hines et al., 2000; Marshall & Harangozo, 2000; Marshall, 2002; Bromwich & Fogt, 2004; Greatbatch & Rong, 2006; Hertzog et al., 2006; Bromwich et al., 2007; Lindsay et al., 2014). For the purposes of our analysis, global fields of daily mean 500 hPa geopotential height (Z_g^{500hPa}), zonal winds at 850 hPa and 200 hPa ($u_{850\ \mathrm{hPa}}$ and $u_{200\ \mathrm{hPa}}$), MSLP, and surface zonal and meridional winds (u_{sfc} and v_{sfc}) are obtained on the provided $2.5^\circ \times 2.5^\circ$ latitude-longitude grid. Daily mean top-of-atmosphere outgoing longwave radiation (OLR) fields are provided on a T62 Gaussian grid and are subsequently regridded to a $2.5^\circ \times 2.5^\circ$ latitude-longitude grid using a bilinear interpolation scheme. To compute indices of tropical variability based on SST data for NNR1, we use version 1.1 of the HadISST SST dataset (Rayner et al.,

2003), which provides monthly global SST on a 1°×1° latitude-longitude grid from 1870 to present.

The JRA-55 reanalysis (Kobayashi et al., 2015), covering the period from 1958 to present, is a more recent atmospheric reanalysis product that aims to take advantage of ongoing improvements in forecasting systems and available observations. As for the NNR1 reanalysis, a frozen analysis system is employed and atmospheric and surface observations are assimilated. The assimilation system used for JRA-55 employs a TL319 resolution operational system with 60 vertical levels. The use of a higher resolution model, together with other updates to the system, has been found to yield improvements in the representation of the synoptic scale atmospheric circulation compared to the previous generation JRA-25 reanalysis (Onogi et al., 2007), although there remain known issues (Harada et al., 2016). Daily mean Z_g^{500hPa}, $u_{850\ \mathrm{hPa}}$, $u_{250\ \mathrm{hPa}}$, u_{sfc}, v_{sfc}, MSLP, and OLR fields are obtained on a 1.25°×1.25° latitude-longitude grid. For SST fields, the model surface brightness temperature provided on a 1.25°×1.25° latitude-longitude grid is used. Where required by the definition of the index as noted below, we regrid the initial fields to a 2.5° × 2.5° latitude-longitude grid using a bilinear interpolation method.

We also examine climate model simulations of the recent past considering a subset of the data submitted to the Coupled Model Intercomparison Project version 5 (CMIP5) (Taylor et al., 2012) which was extensively used to inform the IPCC Fifth Assessment report (AR5) (IPCC, 2013). One of the specific aims of CMIP5 was to promote *a standard set of model simulations in order to ... evaluate how realistic the models are in simulating the recent past* hence providing an ideal dataset for the purposes outlined here namely, to apply Bayesian statistics to reveal differences in the causal relationships between climate teleconnections that occur in similarly forced models as a first step in revealing the dynamical mechanisms responsible for these model differences. While there are examples of constraint based methods applied largely to surface data e.g., (Runge et al., 2019; Donges et al., 2009a), this study is, to our knowledge, the first time a score-based approach has been applied to consider comprehensively the major climate teleconnections and to evaluate climate model performance relative to reanalysis products (here we use the term "score-based" to refer to the use of the posterior probability of a given network and its associated parameters as a measure of model fitness, as described below).

2.1 Indices

To obtain a tractable system for analysis, we characterize the dynamical behavior of the Earth's climate system in terms of a set of process-oriented diagnostics that effectively reduce the volumes of climate model and observational data to a small number of timeseries. These timeseries represent distinct teleconnections within the climate system, i.e., recurrent large-scale modes of variability. Here we consider the major atmospheric synoptic scale teleconnection patterns i.e., the Arctic Oscillation (AO) (Thompson & Wallace, 1998), the phases of the North Atlantic Oscillation (NAO+, NAO-) (Walker, 1923; van Loon & Rogers, 1978), the Pacific North American (PNA) (Wallace & Gutzler, 1981; Horel & Wallace, 1981; Barnston & Livezey, 1987), the two component modes of the Pacific South American pattern (PSA1, PSA2) (Mo & Ghil, 1987; Lau et al., 1994; O'Kane et al., 2017), and the Southern Annular Mode (SAM) (Rogers & van Loon, 1982; Thompson & Wallace, 2000). The tropical Pacific and Indian oceans are the major influences on interannual timescales via the El Niño Southern Oscillation (ENSO) (Walker, 1924; Bjerknes, 1969) and through the Indian Ocean Dipole (IOD) (Saji et al., 1999) whose phases and combined interactions with the extra-tropical synoptic scale atmospheric teleconnections has important impacts on global weather and climate (Geng et al., 2023). At intraseasonal timescales the Madden-Julian Oscillation (Madden & Julian, 1971; Kitsios et al., 2019) is the major mode of tropical convection. Thus, these empirical climate indices provide physically observed modes that allow for an intuitively better understanding of the properties and interactions between the climate modes while reducing the dimensionality of the problem as required for tractable inference.

Following Harries and O'Kane (2021), the monthly time series of the selected climate indices are computed from full gridded fields for the period 1 January 1960 to 30 November 2005, and it is these indices that are used for fitting the DBNs with the time period chosen specifically to facilitate comparison with historical model simulations. We further estimate and remove a linear temporal trend for every index whose respective time series are then standardized to have zero mean and unit variance over the fitting period. These then form a robust and consistent set of thirteen teleconnection indices to then form the random variables (i.e., nodes) of the fitted graphical models. In an alternate fully data-driven approach, the indices would be automatically determined by using community detection methods (Steinhaeuser et al., 2011; Bello et al., 2015) thereby accounting for systematic model biases in the representation of the spatial structures of

the given modes (see for example Kretschmer et al. (2017)). As our purpose here is to infer model biases relative to reference reanalyses, we use the fixed, commonly accepted empirical definitions for the climate indices. The set of indices chosen spans climate variability from intraseasonal through to interannual time-scales. Anomalies are calculated as differences from daily or monthly climatological values with respect to the reference period 1 January 1979 to 30 December 2001. For the reanalyses the calculations of the indices have been validated against publicly available data, including verifying that the spatial patterns of the corresponding indices are in good agreement with previous studies.

As a measure of tropical Pacific ocean variability, we include an updated version of the multivariate ENSO index (MEI) (Wolter & Timlin, 1993, 1998, 2011) defined by Zhang et al. (2019). To characterize activity in the tropical Indian ocean we use the dipole mode index (DMI) (Saji et al., 1999) for the reanalyses and for the CMIP5 models an empirical orthogonal function (EOF) (Lorenz, 1956) based Indian Ocean Dipole (IOD) index to account for possible spatial biases due to shifts in the centres of action that might impact the simulated variability across models. It should be noted that the EOF based IOD and spatially fixed DMI are equivalent in the reanalyses. For tropical variability associated with the Walker circulation and convection over the maritime continent i.e., the Madden-Julian oscillation (MJO), we use the rotated EOF index of Wheeler and Hendon (2004), denoted as RMM1 and RMM2 and defined as the monthly mean of the corresponding daily index. For both the MEI and the RMM1 and RMM2 indices, all of the input fields are evaluated on a common $2.5^\circ \times 2.5^\circ$ latitude-longitude grid.

For the Northern Hemisphere (NH) extratropical atmosphere we include the AO, PNA, and four modes associated with the North Atlantic oscillation and blocking. We define the AO as the leading EOF of monthly mean Z_g^{500hPa} anomalies poleward of 20°N weighted by the square root of the cosine of the gridpoint latitude. The PNA pattern is calculated as the leading mode obtained after performing a VARIMAX rotation (Kaiser, 1958) of the first 10 EOF modes of monthly-standardized monthly mean Z_g^{500hPa} anomalies polewards of 20°N during boreal winter i.e., December, January and February (DJF). The PNA index is then the projection of the standardized height anomalies onto the resulting pattern, standardized by the monthly mean and standard deviation within the climatology reference period. The four Euro-Atlantic circulation regimes are calculated following the approach of Straus et al. (2017) via a k-means clustering analysis of the

Table 1. Climate indices

Climate indices and their geographic distribution		
Tropical	NH	SH
MEI: multivariate ENSO index ENSO: El Niño Southern Oscillation IOD: Indian Ocean Dipole RMM(1&2): Real-time Multivariate MJO index MJO: Madden-Julian Oscillation	AO: Arctic Oscillation NAO(+,-): North Atlantic Oscillation PNA: Pacific-North American pattern AR: Atlantic Ridge SCAND: Scandinavian blocking	SAM: Southern Annular Mode PSA(1&2): Pacific-South American pattern

leading 24 principal components (PCs) of boreal winter anomalies in daily mean Z_g^{500hPa} in the sector [20°N - 80°N, 90°W - 30°E], after applying a 10 day running mean smoothing. The four cluster patterns obtained correspond to the positive and negative NAO phases, NAO^+ and NAO^-, as well as Atlantic Ridge (AR) and Scandinavian blocking (SCAND) patterns, which are associated with blocking events in the Atlantic and western Europe, respectively. The index for each cluster is obtained by projecting the daily or monthly height anomalies onto the anomaly composite associated with one of the four cluster centroids. Once again, each index is standardized using the appropriate monthly mean and standard deviation of the monthly index over the reference period.

For the Southern Hemisphere, the SAM is taken to be the leading EOF of monthly mean Z_g^{500hPa} anomalies poleward of 20°S, where the anomalies are once again weighted by the square root of the cosine of the gridpoint latitude and normalized by the standard deviation of their associated monthly leading PC. The PSA1 and PSA2, are defined as the second and third modes in an EOF analysis of year-round anomalies of daily mean Z_g^{500hPa} polewards of 20°S whose eigenvalues are nearly degenerate indicative of a slowing propagating mode (O'Kane et al., 2017). The respective PSA1 & 2 indices are calculated by projecting Z_g^{500hPa} anomalies onto each mode and normalizing by the standard deviation of the correponding PC over the reference period.

In table 1 we list all indices forming nodes of the DAGs and their geographic locations. We have to also be cognizant that the use of monthly mean data for atmospheric processes, in conjunction with the exclusion of contemporaneous edges in the DAGs as described below, may lead to edges that do not reflect direct physical processes but are instead due to unmodeled subgrid scale interactions taking place on time-scales of less than one month. Our choice to use data with monthly temporal resolution was a pragmatic choice to reduce the computational cost of fitting the models.

3 Structure learning, conditional densities and prior distributions

In applying Bayesian methods for the purpose of comparing climate models and reanalyses the aim is to obtain a sample from the posterior distribution of possible graphical structures. From these posterior samples, we then compute summary statistics including the estimated posterior probability of the existence of one or more edges $\hat{\pi}$, or posterior means $\hat{\beta}$ for parameters conditional on the maximum a posteriori (MAP) structure over individual graphs in the sample. Here we focus solely on summary statistics computed over the full sample. The networks derived from the two reanalysis products provide a baseline to compare individual free-running CMIP5 model simulations. The common historical period chosen for the comparison is the most recent period where satellite data from the atmosphere and surface ocean is sufficiently dense to provide the required temporal resolution and spatial homogeneity to render a reliable reanalysis of the earth system domains. In this section we define the reduced-order models used and provide a brief description of the structure learning approach used; further details are given in Harries and O'Kane (2021).

3.1 Graphical models for teleconnections

To represent the relationships between the selected teleconnections, we consider a class of linear vector autoregressive models (VAR) models that may be formulated as time-homogeneous DBNs. The value of each index i ($i = 1, \ldots, n$) at a given time t, Y_t^i, is treated as a random variable, which is graphically represented by a node in the model DAG. As many of the variables of interest, and particularly those in the tropics associated with seasonal to interannual variability, exhibit substantial autocorrelation, this set of random variables is expanded to include the lagged values of the indices $Y_{t-\tau}^i$ at previous times $t-\tau$ (Kjærulff, 1995; Friedman et al., 1998; K. Murphy & Mian, 1999; K. P. Murphy & Russell, 2002), up to some maximum lag $\tau_{\max}$. The DBNs reported here based on the CMIP5 and reanalysis data have been restricted to a maximum time lag of 6 months.

Dependence of the current state Y_t^i on the past value of an index $Y_{t-\tau}^j$ is represented graphically as a directed edge from the node representing $Y_{t-\tau}^j$ to the node for Y_t^i (Eichler, 2012; Runge, 2018). The time-ordered nature of the interactions requires that edges only point from past to present. In addition, we exclude the possibility of contemporaneous

dependencies among variables. Doing so implies that the resulting models satisfy structural modularity (Friedman & Koller, 2003), which simplifies model fitting. However, interactions that occur on time-scales shorter than the data sampling frequency (in this case, monthly) cannot be accounted for and may, where said interactions at lags of less than one time-step are relevant, result in model misspecification. The set of lagged values that are assumed to influence the present state of an index Y_t^i, satisfying these constraints, are referred to as the parents $\mathrm{pa}_G(Y_t^i)$ of Y_t^i, with Y_t^i being the child node. Graphically, specification of the parents for each index determines the structure of the corresponding DAG by defining which directed edges are present in the graph.

In reality, in a multiscale climate system, one cannot assume temporal-homogeneity of the interactions between modes. The various modes of variability are known to interact across spatio-temporal scales. An example is the interaction between synoptic variability of persistent coherent states in the South Pacific mid-troposphere represented by the subseasonal PSA1, with the atmospheric response to convection over the maritime continent represented by the intra-seasonal MJO and interannual variations of tropical Pacific sea surface temperatures i.e., ENSO. Non-stationary forcing (i.e., anthropogenic warming) over the data period may similarly manifest as changes in the graph structure or associated model parameters. Where secular trends and regimes are present one must be aware of the possibility that either the graph structure, parameters or both are simultaneously dynamic over time (Wu et al., 2018; Saggioro et al., 2020), thereby dramatically increasing the computational task when employing score-based methods for structure learning. Our focus here is on model uncertainty and biases hence we restrict ourselves to the case of homogeneous models.

We assume that, conditional on the values of its parents $\mathrm{pa}_G(Y_t^i)$, each index Y_t^i is normally distributed with mean μ_t^i given by a linear function of the parent variable values (Punskaya et al., 2002; Lèbre et al., 2010),

$$\begin{aligned} Y_t^i|\mathrm{pa}_G(Y_t^i), \tilde{\tau}_i^2 &\sim N(\mu_t^i, \tilde{\tau}_i^{-2}), \\ \mu_t^i &= \beta_0^i + \sum_{j=1}^{p_i} \beta^i_{(k_j,\tau_j)} Y^{k_j}_{t-\tau_j}, \end{aligned} \tag{1}$$

where p_j denotes the size of the set of parents i.e., $pa_G(Y_t^i) = \{Y^{k_j}_{t-\tau_j} | j = 1, \ldots, p_i\}$, k_j the index of the j^{th} member of the parent set, and τ_j the corresponding time lag. This is a specialization of the BGe model (Geiger & Heckerman, 1994). With these assump-

tions, the model likelihood can then be evaluated by replicating the graph structure over the full time series.

Fitting this model also requires that priors be specified for the regression coefficients $\beta^i_{(k_j,\tau_j)}$ corresponding to edges in the parent set and the conditional precision $\tilde{\tau}_i^2$. We choose to use conjugate normal-gamma priors

$$\begin{aligned} \tilde{\tau}_i^2 &\sim \text{Gamma}(a_\tau, b_\tau), \\ \beta_0^i|\tilde{\tau}_i^2 &\sim N\left(0, \frac{\nu_i^2}{\tilde{\tau}_i^2}\right), \\ \beta^i_{(k_j,\tau_j)}|\tilde{\tau}_i^2, \text{pa}_G(Y_t^i) &\sim N\left(0, \frac{\nu_i^2}{\tilde{\tau}_i^2}\right). \end{aligned} \tag{2}$$

where a_τ, b_τ, and ν_i^2 are prior hyperparameters. Note that, in principle, to define the full posterior distribution for the model structure and parameters it would also be necessary to choose a set of pseudo-priors for those $\beta^i_{(k_j,\tau_j)}$ corresponding to edges not present in the DAG. However, these pseudo-priors do not enter into the implementation of the particular sampling scheme used below (Godsill, 2001), and hence we leave them unspecified. Alternative choices for the hyperparameters a_τ, b_τ, and ν_i^2 allow varying levels of regularization to be imposed. For consistency, in this study the hyperparameters were chosen to match those previously used in analysing the reanalysis datasets, i.e., $a_\tau = 1.5$, $b_\tau = 20$, and $\nu_i^2 = 3$, for $i = 1, \ldots, n$. The unconditional prior distribution for a given coefficient β following from Eq. (2), after marginalizing out the precision $\tilde{\tau}_i^2$, is a generalized t-distribution with zero mean, scale parameter $\nu_i^2/(a_\tau b_\tau)$, and $2a_\tau$ degrees of freedom. Hence, this choice of hyperparameters corresponds to a generalized t-prior for the β parameters, with a 95% prior highest density interval (HDI) of $-1 \leq \beta \leq 1$. This choice leads to somewhat informative priors, but qualitatively similar results were found for the reanalyses data using much more weakly informative choices of $a_\tau = 0.5$, $b_\tau = 10$, $\nu_i^2 \approx 2$ (corresponding to a 90% prior HDI of $-4 \leq \beta \leq 4$ and prior 1% and 99% percentiles for τ^2 of 7.6×10^{-4} and 33.2, respectively).

3.2 Structure learning

Models in the class described above are fully specified by the parent sets for each index, which define a graph G, and the corresponding collection of parameters θ consisting of the coefficients $\beta^i_{(k_j,\tau_j)}$ and the conditional precision τ_i^2. When the graph structure (i.e., the parent sets) is not pre-specified, fitting the homogeneous DBN requires learning both the structure G and the parameters θ. Given data $D = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$, where

$\boldsymbol{y}_t$ denotes the values of the random variables $\boldsymbol{Y}_t = (Y_t^1, \ldots, Y_t^n)^T$ at time t, learning the structure G and parameters θ can be done in two steps, since

$$P(\theta, G|D) = P(\theta|G, D)P(G|D). \tag{3}$$

In the first, structure learning step, the posterior distribution over possible structures $P(G|D)$ is determined. For a given choice of G, the corresponding posterior distribution for the parameters $P(\theta|G, D)$ may be then be computed.

Here we take a score-based approach where the graph G is estimated based on maximizing a suitable score function (Cooper & Herskovits, 1992; Geiger & Heckerman, 1994; Heckerman et al., 1995), in our case the marginal likelihood $P(D|G)$. Rather than finding a single optimal model, we attempt to account for model uncertainty by sampling from the full posterior distribution of possible graphs $P(G|D)$ (Madigan et al., 1995). Sampling the posterior $P(G|D)$ requires evaluation of the marginal likelihood

$$P(D|G) = \int d\theta P(D|G, \theta)P(\theta|G), \tag{4}$$

where $P(\theta|G)$ denotes a set of priors for the full set of parameters θ conditional on the structure of the graph, and we have used the shorthand $\int d\theta$ to denote marginalization. The factor $P(D|G, \theta)$ is simply the likelihood under the model. For the DBN models described above, the marginalization in $P(D|G)$ can be evaluated analytically, and hence it is possible to sample the posterior distribution $P(G|D)$ using the MC^3 scheme of Madigan et al. (1995). Further details on the Markov chain Monte Carlo (MCMC) methods used are given in Appendix A. Briefly, given a current candidate structure G, the sampling scheme proceeds by proposing a new structure G' according to a proposal distribution $q_G(G'; G)$. The proposal G' is accepted with probability

$$\alpha = \min\left\{1, \frac{q_G(G; G')}{q_G(G'; G)} \frac{P(D|G')}{P(D|G)} \frac{P(G')}{P(G)}\right\}; \tag{5}$$

otherwise, the current state G is retained. In Eq. (5), $P(G)$ denotes the prior distribution for the structure G. We choose structurally modular priors of the form

$$P(G) = \prod_{i=1}^{n} P(\mathrm{pa}_G(Y_t^i)) \tag{6}$$

such that $P(G)$ factorizes into a product over priors on the parent sets. We adopt uniform priors over the set of parent sets for each index, subject to the constraint that the maximum time-lag is $\tau_{\max} = 6$ months. We further also impose a maximum size $p_{\max} =$

10 on the allowed parent sets in order to sparsify the networks. With these constraints,

$$P(\mathrm{pa}_G(Y_t^i)) = \begin{cases} \left[\sum_{j=0}^{p_{\max}} \binom{n\tau_{\max}}{j}\right]^{-1}, & |\mathrm{pa}_G(Y_t^i)| \leq p_{\max}, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

We also adopt a uniform proposal density on graphs G' in the neighborhood of the current graph G,

$$q_G(G';G) = \begin{cases} \frac{1}{|\mathrm{nhd}(G)|}, & G' \in \mathrm{nhd}(G), \\ 0, & \text{otherwise.} \end{cases}$$

The neighborhood $\mathrm{nhd}(G)$ of a graph G consists of the set of graphs that can be reached from that structure by a single move in a predefined move set. The possible moves that we allow include addition of a single edge, deletion of a single edge, or an exchange of two edges (Grzegorczyk & Husmeier, 2011). The neighborhood of a graph contains only those graphs that can be reached by performing one of these three moves, subject to the imposed condition on the maximum parent set size. Inclusion of the exchange move allows slightly more efficient exploration of the model space; for the sampler settings used, overall acceptance rates ranging from $0.16-0.37$ are obtained, dependent on the particular index.

Our DBN framework uses the simplest case of a linear model with conjugate priors on the parameters defining the conditional PDFs, together with priors on the structures to ensure structural modularity. No prior restriction has been enforced to ensure stationarity of the resulting autoregressive model. Additionally, no attempt has been made to incorporate pre-existing or expert knowledge into the definition of the chosen priors.We note in particular that, in practice, it may be more appropriate to utilize more informative priors that incorporate such information. By doing so, posterior inferences may be regularized so as to obtain more reliable estimates, given the generally limited sample size available for historical observations. This does, however, in general prevent analytical evaluation of the relevant posterior distributions as is possible with the choice of conjugate priors used here.

With the use of structurally modular priors and exclusion of contemporaneous edges, the full posterior distribution for the structure G factorizes into separate terms for each index, and hence sampling can be done in parallel for each index separately. For each index, posterior samples were obtained by running 8 chains of length 1×10^7 samples, discarding the first 250,000 samples as burn-in. Chain convergence was assessed by con-

sidering the homogeneity of the distribution of parent sets within chains using χ^2 and Kolmogorov-Smirnov tests (Brooks et al., 2003) for each index, in addition to trace plots for individual edge indicators. We found some evidence of non-homogeneity across chains based on the full sample, suggesting further sampling may be required for convergence, although we expect the posterior estimates obtained to be sufficient for the qualitative comparisons reported here. Various choices of thinning parameter were considered to determine the number of retained samples based on convergence rates. Qualitatively our finding was that the evaluated graphs were insensitive to thinning up to a factor of 100; results presented here are based on the full set of posterior samples.

3.3 Posterior summaries

From a sample of size S from the posterior distribution $P(G|D)$, distributional estimates for derived quantities of interest Δ may be obtained by averaging over the sample (Madigan & Raftery, 1994; Draper, 1995),

$$\Pr(\Delta|D) = \sum_{G\in\mathcal{G}} \Pr(\Delta|G,D)P(G|D) \approx \frac{1}{S}\sum_{s=1}^{S} \Pr(\Delta|G^{(s)},D), \tag{8}$$

where $G^{(s)}$ is the s^{th} structure sample. In particular, structural uncertainties may be quantified by taking Δ to be an indicator function for the presence of a given edge, with Eq. (8) quantifying the posterior probability $\hat{\pi}$ for the presence of that edge, given the observed data. In the results to follow, we display the estimates $\hat{\pi}$ found for edges in the reanalysis datasets and the CMIP5 models considered. As argued in the introduction, differences between models where the presence of an edge in one model is supported with high posterior probability may be indicative of important model biases. This motivates comparing the (marginal) posterior distributions for the individual edges between the CMIP5 models and reanalyses. Below, we present graphical summaries of these estimated posterior weights $\hat{\pi}$ by showing each corresponding edge with width proportional to $\hat{\pi}$; for clarity, only those edges with $\hat{\pi} > 0.5$ are shown in figures. For qualitative comparisons, it is useful to have a heuristic measure of the overall difference in the distribution of posterior mass for possible edges. For visualization purposes, we sort models according to an earthmover's or Wasserstein distance (Villani, 2009), computed between histograms of the posterior weights for all edges simultaneously, for each reanalyis and CMIP5 model. Utilizing JRA-55 as an initial reference, models are shown ordered such that the distances between adjacent models are minimized. We also repeat these calculations alternatively

using the sum of the pairwise Wasserstein distances between weights for individual edges or the sum of the individual Kullback-Leibler divergences (Hall, 1987; Burnham & Anderson, 2002) between the edge posterior distributions. When computed either for all possible edges, or for particular subsets of nodes, this provides a useful means of qualitatively assessing similarities between the models and reanalyses (computing a genuine divergence between the joint posterior distributions is not feasible with the available sampling).

It should be emphasized that the graphs presented in the following section do not correspond to a single DBN model, but are rather summaries of the presence/absence of an edge over the full sample. To compare the sign and magnitude of the association between indices requires conditioning on a particular DBN, defining a particular set of regression coefficients $\beta^i_{(k_j,\tau_j)}$. In particular, we may consider the DBN corresponding to the posterior mode, or maximum a posteriori (MAP) estimate. As we have chosen conjugate normal-gamma priors for the model parameters, it is straightforward to evaluate the posterior distributions for the coefficients $\beta^i_{(k_j,\tau_j)}$ analytically, and hence obtain posterior means and 95% HDIs.

4 Results

We now focus on the DAG edge weights, more specifically the summary representations of the estimated posterior probabilities for edges between nodes. Major differences between the DAGs of the CMIP5 models and those of the reanalyses are assumed to be indicative of systematic biases in the models' representation of the selected internal modes of variability, their teleconnections and interactions. While we will discuss differences between individual CMIP5 models and observations as represented in the reanalyses, our objective is to demonstrate the utility of the approach to identify model error in a mathematically consistent and justifiable Bayesian framework that also provides easy physical interpretability. For example, while an univariate autoregressive analysis of a given climate index may be applied to determine autocorrelation, it can only provide a quantitative assessment of the timescale of the autocorrelation. In the results that we present next, we show how the DBNs reveal not simply information on autocorrelation but also any lagged relationships that might be the potential cause of biases associated with either too weak or strong influence of a given mode on the wider climate system.

We summarize the results of sampling in a variety of ways. Figure 1 shows three alternative representations of the graphs associated with tropical variability in the JRA-55 reanalysis. Panel (a) shows a DAG representation of the estimated posterior edge probabilities for the tropical ENSO (MEI), MJO (RMM1 & 2) and IOD modes. Here the "child" is the node associated with a given index at $t = 0$, whose "parents" are any node for a given index at lags $t = 1, \ldots, 6$ months for edges with an estimated posterior weight greater than 0.5. Coefficients linking any parent to any other parent are not allowed nor is the present allowed to influence the past. In panel (c) a reduced representation of the same DAG that appears in panel (a) is shown where it is assumed the edge exists only between parent and child but the implied autocorrelation is now shown by the arrows. Panel (b) shows the same information without thresholding of the posterior weights as a heat map. Here the rows indicate the index at time $t = 0$ (i.e., the child node) whereas the columns show the parents for each child at lags up to 6 months. Each row is calculated from the retained sample of possible graphs from the posterior distribution, after discarding burn-in samples. The shading indicates the value of the posterior weight corresponding to the probability that an edge exists between parent and child.

4.1 Full year networks for monthly indices

In figures 2 & 3 we show the network summaries for the tropical modes for both JRA-55 and NNR1 reanalyses and seven CMIP5 models conditioned over the available timeseries data. Here we can see the longest autocorrelations occur for ENSO and the MJO at lags of up to 4 months for the MJO and across the considered 6 months lags for ENSO. This is completely consistent with our current understanding of the associated dynamics and predictability of ENSO (O'Kane et al., 2020) and the MJO (Kitsios et al., 2019) on seasonal timescales. The major differences between the reanalyses concern the strong relationship in NNR1 between midlatitude blocking associated with the Atlantic Ridge index and convection in the Indian and maritime continents i.e., the IOD and MJO respectively relative to a much weaker teleconnection present in JRA-55. There are also differences in amplitude for posterior weights indicating associations between the high latitude AO and SAM modes and the PSA1 to the MJO. In spite of these differences, there is an obvious consistency between the two reanalyses even though they have been generated using contrasting data assimilation schemes, model resolutions and configu-

rations with their development displaced in time by well over a decade. Harries and O'Kane (2021) provide an extensive comparison and discussion of the two reanalysis products.

We next turn our focus onto the performance of the selected CMIP5 models. There are immediate commonalities observable across the CMIP5 models with regard to the autocorrelation present in each of the DAGs. Apart from MIROC5, all models exhibit longer autocorrelation in their IOD than are present in the reanalyses. MIROC5, GFDL-ESM2M and to a lesser degree ACCESS1-0 have a significantly shorter ENSO autocorrelation than observed. ACCESS1-0 has a clear lagged influence on the MEI that is not reproduced in any of the other models considered. All models, with the exception of MIROC5, have autocorrelation in the component modes of the MJO and at increased lags of up to 6 months for GFDL-ESM2M. Due to their long autocorrelation times and memory, the tropical modes can exert sustained influence on the purely atmospheric modes, however the synoptic timescales of the atmospheric modes with short autocorrelation beyond a few weeks to a month serves as a sign that we should not expect a strong influence of the atmospheric modes on the tropics at longer time lags which provides a physical basis for interpreting the sparsified graphs.

In table 2 we show an ordering of models referenced to JRA-55 in terms of the heuristic measures of similarity based on either the earthmover's distance or Kullback-Leibler divergence; models are sorted from smallest to largest distance when calculated over the tropics, both hemispheres and for teleconnections between the respective hemispheres and the tropics. For the Kullback-Leibler divergence, rather than considering divergences based on the the prior and posterior distributions, we show the divergence based on distributions induced by the NNR1 and CMIP models with respect to the reference distribution from the JRA55 reanalysis. Results were also found to be largely insensitive to thinning of the samples by factors of 100 and 1000 thereby reducing any dependencies present between successive samples. In broad terms and regardless of metric, the ordering reflects what is expected i.e., that the reanalyses JRA and NNR1 are always paired together, and the CMIP models vary in order dependent on model, the particular subset of indices being looked at, and by geographic region. In table 3 the model ordering is determined such that the distance between adjacent models is minimized rather than simply ranking by the distance to the reference model. We use this ordering in the heatmaps that follow so that adjacent models are as similar as possible. In both tables the values of the distance of each model referenced to JRA-55 is shown in the bracketed values. In

particular where the ordering is based on the pairwise Wasserstein distance or Kulback-Leibler divergence it is apparent that the considered CMIP5 models form a quite distinct class with respect to the reanalyses.

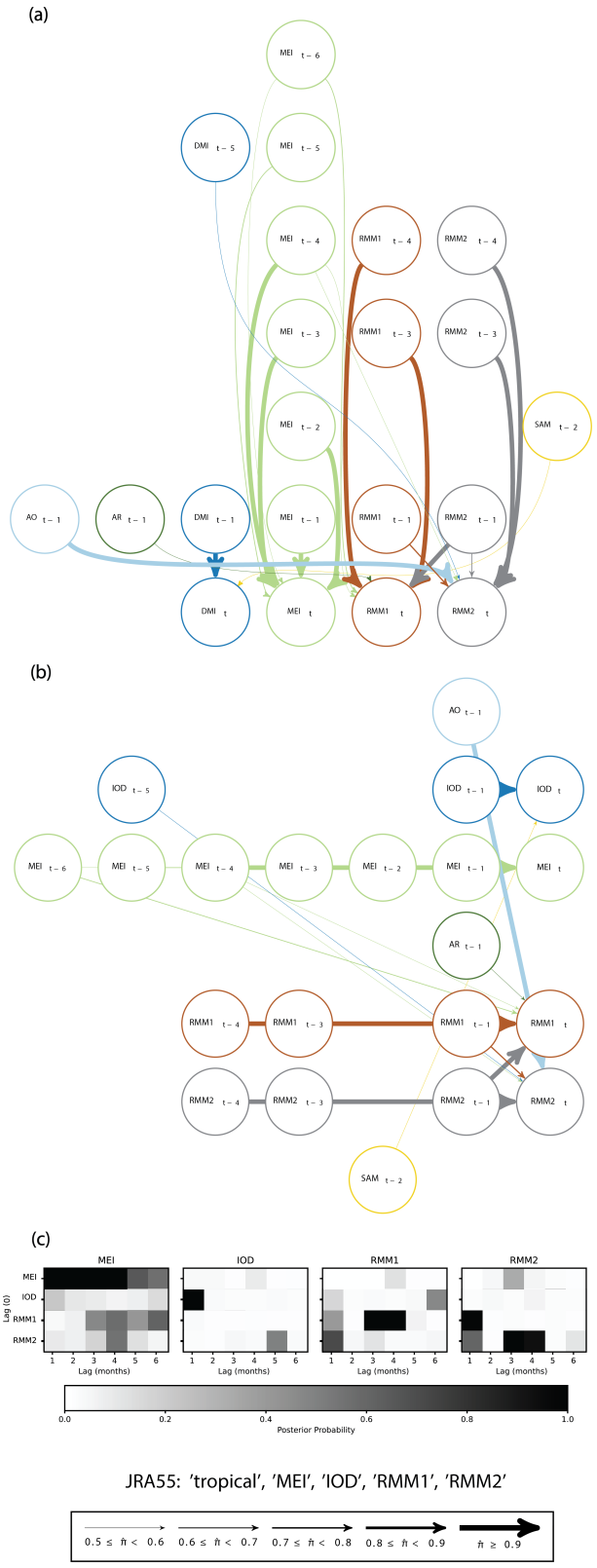
(a)
MEI t − 6
DMI t − 5
MEI t − 5
MEI t − 4
RMM1 t − 4
RMM2 t − 4
MEI t − 3
RMM1 t − 3
RMM2 t − 3
MEI t − 2
SAM t − 2
AO t − 1
AR t − 1
DMI t − 1
MEI t − 1
RMM1 t − 1
RMM2 t − 1
DMI t
MEI t
RMM1 t
RMM2 t
(b)
AO t − 1
IOD t − 5
IOD t − 1
IOD t
MEI t − 6
MEI t − 5
MEI t − 4
MEI t − 3
MEI t − 2
MEI t − 1
MEI t
AR t − 1
RMM1 t − 4
RMM1 t − 3
RMM1 t − 1
RMM1 t
RMM2 t − 4
RMM2 t − 3
RMM2 t − 1
RMM2 t
SAM t − 2
(c)
MEI
IOD
RMM1
RMM2
Lag (0)
Lag (months)
Posterior Probability
0.0
0.2
0.4
0.6
0.8
1.0
JRA55: 'tropical', 'MEI', 'IOD', 'RMM1', 'RMM2'
0.5 ≤ π < 0.6
0.6 ≤ π < 0.7
0.7 ≤ π < 0.8
0.8 ≤ π < 0.9
π ≥ 0.9

Figure 1.

Figure 1. (Previous page.) Alternative representations of the estimated posterior edge probabilities for the tropical climate modes as calculated from the JRA-55 reference data. We show graphs where the "child" is the node associated with a given index at $t = 0$, whose "parents" are any node for a given index at lags $t = 1, \ldots, 6$ months for edges with an estimated posterior weight greater than 0.5. Panel (a) shows the format used in Harries and O'Kane (2021). In panel (b) a reduced representation of the same DAG is shown where it is assumed the edge exists only between parent and child. Panel (c) shows the same information without thresholding of the posterior weights as a "heat map". Here the rows represent the "children" with the columns showing the "parents". The shading indicates the value of the posterior weight.

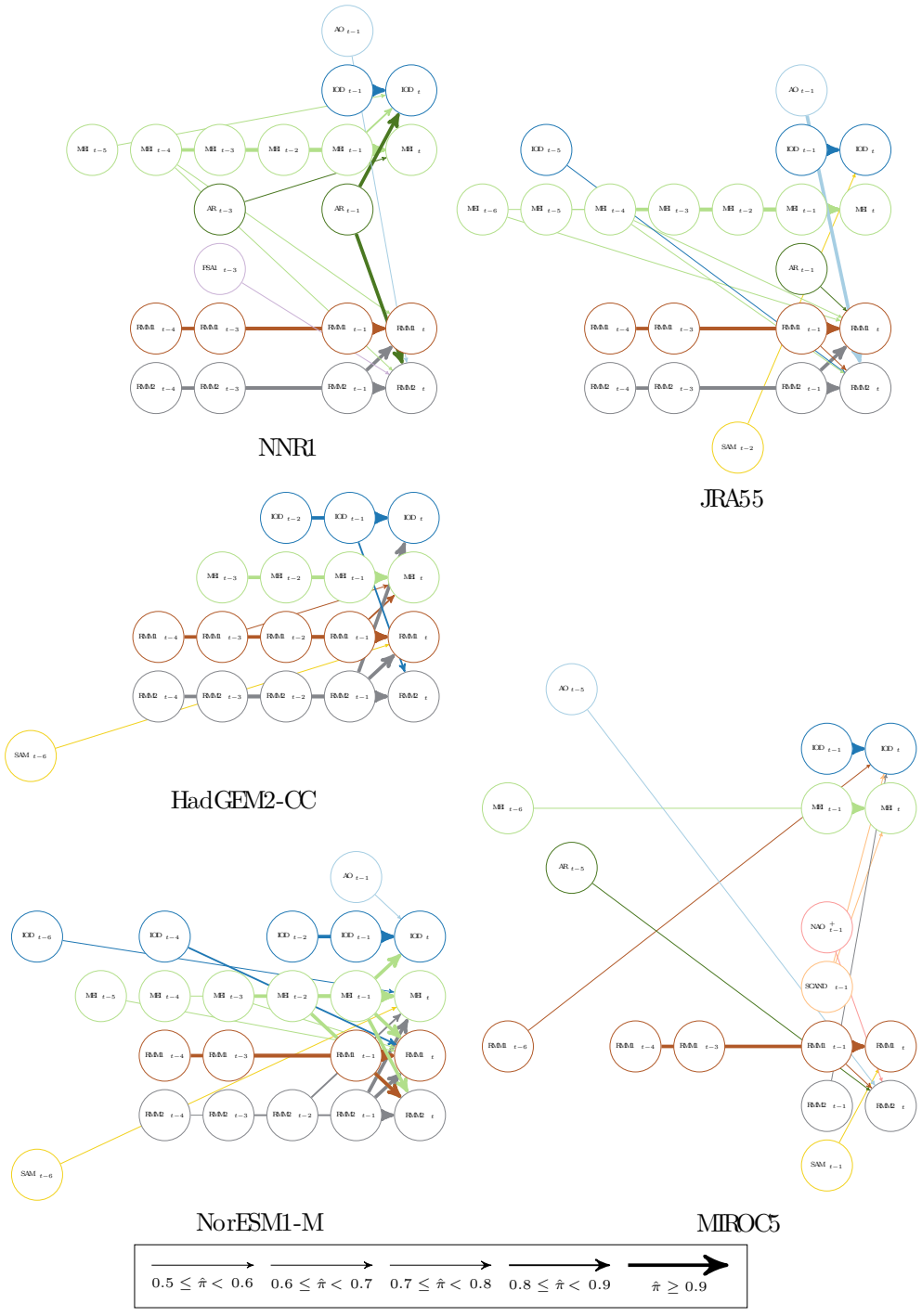

Figure 2. Edge posterior probabilties for the tropical indices calculated over all seasons (ALL: 'tropical') for the JRA-55 and NNR1 reanalyses and for the HadGEM-CC, NorESM1-M, and MIROC5 historical CMIP5 model simulations. Only edges with an estimated posterior weight greater than 0.5 are shown.

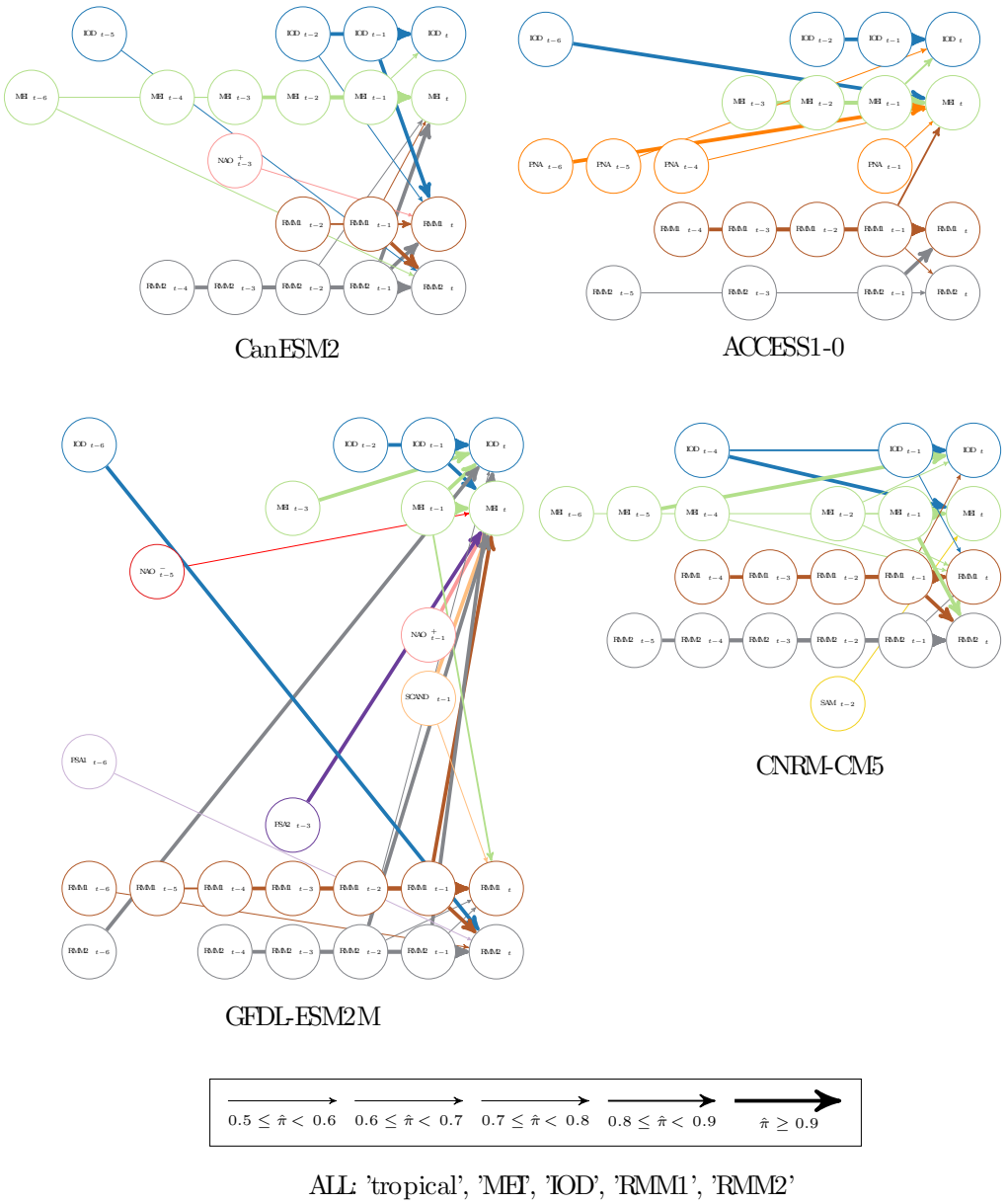

Figure 3. As for figure 2 but for CanESM2, ACCESS1-0, GFDL-ESM2M, and CNRM-CM5.

Table 2. Wasserstein distance and Kullback-Leibler divergences. Calculations are over posterior weights for indices related to specific regions and describing teleconnections between regions. The distance of each model referenced to JRA-55 is shown in the bracketed values.

Models ordered by geographically determined Wasserstein distance

Region →	All (global)	Tropical	NH	SH	NH-Tropical	SH-Tropical
Order ↓	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)
1	NNR1 (10.0)	NNR1 (5.8)	NNR1 (4.8)	NNR1 (3.2)	NNR1 (5.6)	NNR1 (7.3)
2	HadGEM2-CC (10.9)	ACCESS1-0 (6.0)	ACCESS1-0 (5.3)	HadGEM2-CC (7.5)	HadGEM2-CC (6.7)	HadGEM2-CC (7.6)
3	CNRM-CM5 (16.9)	MIROC5 (6.7)	MIROC5 (6.7)	MIROC5 (7.9)	GFDL-ESM2M (10.9)	CanESM2 (9.1)
4	NorESM1-M (19.0)	GFDL-ESM2M (8.8)	GFDL-ESM2M (7.5)	ACCESS1-0 (8.8)	CNRM-CM5 (11.3)	NorESM1-M (9.7)
5	CanESM2 (19.7)	NorESM1-M (10.0)	CNRM-CM5 (8.2)	CNRM-CM5 (9.6)	CanESM2 (12.3)	CNRM-CM5 (11.9)
6	GFDL-ESM2M (23.0)	HadGEM2-C (10.7)	CanESM2 (8.4)	CanESM2 (10.9)	NorESM1-M (14.2)	GFDL-ESM2M (12.6)
7	MIROC5 (26.4)	CanESM2 (11.1)	HadGEM2-CC (9.3)	NorESM1-M (11.0)	MIROC (15.5)	MIROC5 (12.9)
8	ACCESS1-0 (36.2)	CNRM-CM5 (12.2)	NorESM1-M (9.8)	GFDL-ESM2M (11.6)	ACCESS1-0 (19.2)	ACCESS1-0 (21.5)

Models ordered by geographically determined sum over pairwise Wasserstein distances

Region →	All (global)	Tropical	NH	SH	NH-Tropical	SH-Tropical
Order ↓	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)
1	NNR1 (36.2)	NNR1 (17.0)	NNR1 (11.4)	NNR1 (7.9)	NNR1 (28.4)	NNR1 (24.8)
2	HadGEM2-CC (75.1)	HadGEM2-CC (28.3)	HadGEM2-CC (30.5)	MIROC5 (12.9)	HadGEM2-CC (58.8)	HadGEM2-CC (42.7)
3	ACCESS1-0 (76.0)	CanESM2 (28.6)	ACCESS1-0 (32.9)	CanESM2 (15.6)	ACCESS1-0 (63.1)	ACCESS1-0 (44.6)
4	CanESM2 (80.8)	NorESM1-M (29.3)	NorESM1-M (33.3)	ACCESS1-0 (16.3)	CanESM2 (63.2)	MIROC5 (44.9)
5	NorESM1-M (80.9)	ACCESS1-0 (29.8)	CanESM2 (35.6)	HadGEM2-CC (16.6)	NorESM1-M (64.2)	CanESM2 (45.2)
6	MIROC5 (81.5)	MIROC5 (30.3)	MIROC5 (35.9)	CNRM-CM5 (16.6)	MIROC5 (65.9)	NorESM1-M (48.0)
7	CNRM-CM5 (84.5)	CNRM-CM5 (32.0)	CNRM-CM5 (36.6)	NorESM1-M (17.7)	CNRM-CM5 (68.0)	CNRM-CM5 (48.6)
8	GFDL-ESM2M (96.9)	GFDL-ESM2M (40.2)	GFDL-ESM2M (38.2)	GFDL-ESM2M (18.5)	GFDL-ESM2M (78.4)	GFDL-ESM2M (58.7)

Models ordered by geographically determined Kullback-Leibler divergence

Region →	All (global)	Tropical	NH	SH	NH-Tropical	SH-Tropical
Order ↓	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)
1	NNR1 (30.2)	NNR1 (17.0)	NNR1 (8.0)	NNR1 (5.3)	NNR1 (24.9)	NNR1 (22.3)
2	MIROC5 (132.7)	NorESM1-M (52.4)	GFDL-ESM2M (45.7)	CanESM2 (15.0)	MIROC5 (114.5)	HadGEM2-C (70.4)
3	ACCESS1-0 (146.7)	CanESM2 (55.4)	HadGEM2-CC (47.1)	ACCESS1-0 (18.2)	HadGEM2-CC (125.7)	NorESM1-M (72.0)
4	HadGEM2-CC (153.9)	CNRM-CM5 (67.1)	CanESM2 (55.1)	NorESM1-M (19.5)	NorESM1-M (127.9)	ACCESS1-0 (85.6)
5	NorESM1-M (157.2)	ACCESS1-0 (67.4)	MIROC5 (60.8)	MIROC5 (22.0)	ACCESS1-0 (131.7)	CNRM-CM5 (96.4)
6	CanESM2 (172.6)	MIROC5 (80.1)	NorESM1-M (69.6)	CNRM-CM5 (28.2)	CanESM2 (153.1)	MIROC5 (108.2)
7	CNRM-CM5 (178.2)	HadGEM2-CC (86.6)	ACCESS1-0 (76.3)	HadGEM2-CC (29.3)	CNRM-CM5 (156.2)	CanESM2 (108.6)
8	GFDL-ESM2M (206.2)	GFDL-ESM2M (117.9)	CNRM-CM5 (100.7)	GFDL-ESM2M (33.2)	GFDL-ESM2M (173.0)	GFDL-ESM2M (151.1)

manuscript submitted to *Journal of Advances in Modeling Earth Systems (JAMES)*

Table 3. Wasserstein distance and Kullback-Leibler divergences. Calculations are over posterior weights for indices related to specific regions and describing teleconnections between regions. The distance of each model referenced to JRA-55 is shown in the bracketed values. Here the model ordering was determined such that the distance between adjacent models is minimized rather than simply ranking by the distance to the reference model.

Models ordered by geographically determined Wasserstein distance

Region →	All (global)	Tropical	NH	SH	NH-Tropical	SH-Tropical
Order ↓	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)
1	NNR1 (10.0)	NNR1 (5.8)	NNR1 (4.8)	NNR1 (3.2)	NNR1 (5.6)	NNR1 (7.3)
2	HadGEM2-CC (10.9)	MIROC5 (6.7)	MIROC5 (6.7)	HadGEM2-CC (7.5)	HadGEM2-CC (6.7)	HadGEM2-CC (7.6)
3	CNRM-CM5 (16.9)	HadGEM2-C (10.7)	CNRM-CM5 (8.2)	NorESM1-M (11.0)	CNRM-CM5 (11.3)	CanESM2 (9.1)
4	NorESM1-M (19.0)	CanESM2 (11.1)	HadGEM2-CC (9.3)	ACCESS1-0 (8.8)	GFDL-ESM2M (10.9)	MIROC5 (12.9)
5	GFDL-ESM2M (23.0)	NorESM1-M (10.0)	GFDL-ESM2M (7.5)	CNRM-CM5 (9.6)	CanESM2 (12.3)	NorESM1-M (9.7)
6	MIROC5 (26.4)	ACCESS1-0 (6.0)	ACCESS1-0 (5.3)	GFDL-ESM2M (11.6)	ACCESS1-0 (19.2)	CNRM-CM5 (11.9)
7	CanESM2 (19.7)	GFDL-ESM2M (8.8)	CanESM2 (8.4)	CanESM2 (10.9)	MIROC (15.5)	GFDL-ESM2M (12.6)
8	ACCESS1-0 (36.2)	CNRM-CM5 (12.2)	NorESM1-M (9.8)	MIROC5 (7.9)	NorESM1-M (14.2)	ACCESS1-0 (21.5)

Models ordered by geographically determined sum over pairwise Wasserstein distances

Region →	All (global)	Tropical	NH	SH	NH-Tropical	SH-Tropical
Order ↓	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)
1	NNR1 (36.2)	NNR1 (17.0)	NNR1 (11.4)	NNR1 (7.9)	NNR1 (28.4)	NNR1 (24.8)
2	HadGEM2-CC (75.1)	HadGEM2-CC (28.3)	HadGEM2-CC (30.5)	ACCESS1-0 (16.3)	HadGEM2-CC (58.8)	ACCESS1-0 (44.6)
3	ACCESS1-0 (76.0)	ACCESS1-0 (29.8)	NorESM1-M (33.3)	MIROC5 (12.9)	ACCESS1-0 (63.1)	HadGEM2-CC (42.7)
4	CanESM2 (80.8)	CanESM2 (28.6)	CanESM2 (35.6)	HadGEM2-CC (16.6)	NorESM1-M (64.2)	CanESM2 (45.2)
5	NorESM1-M (80.9)	MIROC5 (30.3)	ACCESS1-0 (32.9)	NorESM1-M (17.7)	CanESM2 (63.2)	NorESM1-M (48.0)
6	MIROC5 (81.5)	NorESM1-M (29.3)	CNRM-CM5 (36.6)	CanESM2 (15.6)	MIROC5 (65.9)	MIROC5 (44.9)
7	CNRM-CM5 (84.5)	CNRM-CM5 (32.0)	MIROC5 (35.9)	CNRM-CM5 (16.6)	CNRM-CM5 (68.0)	CNRM-CM5 (48.6)
8	GFDL-ESM2M (96.9)	GFDL-ESM2M (40.2)	GFDL-ESM2M (38.2)	GFDL-ESM2M (18.5)	GFDL-ESM2M (78.4)	GFDL-ESM2M (58.7)

Models ordered by geographically determined Kullback-Leibler divergence

Region →	All (global)	Tropical	NH	SH	NH-Tropical	SH-Tropical
Order ↓	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)	JRA-55 (0)
1	NNR1 (30.2)	NNR1 (17.0)	NNR1 (8.0)	NNR1 (5.3)	NNR1 (24.9)	NNR1 (22.3)
2	MIROC5 (132.7)	ACCESS1-0 (67.4)	HadGEM2-CC (47.1)	ACCESS1-0 (18.2)	MIROC5 (114.5)	ACCESS1-0 (85.6)
3	CNRM-CM5 (178.2)	HadGEM2-CC (86.6)	NorESM1-M (69.6)	MIROC5 (22.0)	CNRM-CM5 (156.2)	CanESM2 (108.6)
4	ACCESS1-0 (146.7)	CanESM2 (55.4)	ACCESS1-0 (76.3)	CanESM2 (15.0)	ACCESS1-0 (131.7)	HadGEM2-C (70.4)
5	HadGEM2-CC (153.9)	MIROC5 (80.1)	GFDL-ESM2M (45.7)	CNRM-CM5 (28.2)	HadGEM2-CC (125.7)	MIROC5 (108.2)
6	NorESM1-M (157.2)	CNRM-CM5 (67.1)	MIROC5 (60.8)	HadGEM2-CC (29.3)	NorESM1-M (127.9)	CNRM-CM5 (96.4)
7	CanESM2 (172.6)	NorESM1-M (52.4)	CNRM-CM5 (100.7)	NorESM1-M (19.5)	CanESM2 (153.1)	NorESM1-M (72.0)
8	GFDL-ESM2M (206.2)	GFDL-ESM2M (117.9)	CanESM2 (55.1)	GFDL-ESM2M (33.2)	GFDL-ESM2M (173.0)	GFDL-ESM2M (151.1)

manuscript submitted to *Journal of Advances in Modeling Earth Systems (JAMES)*

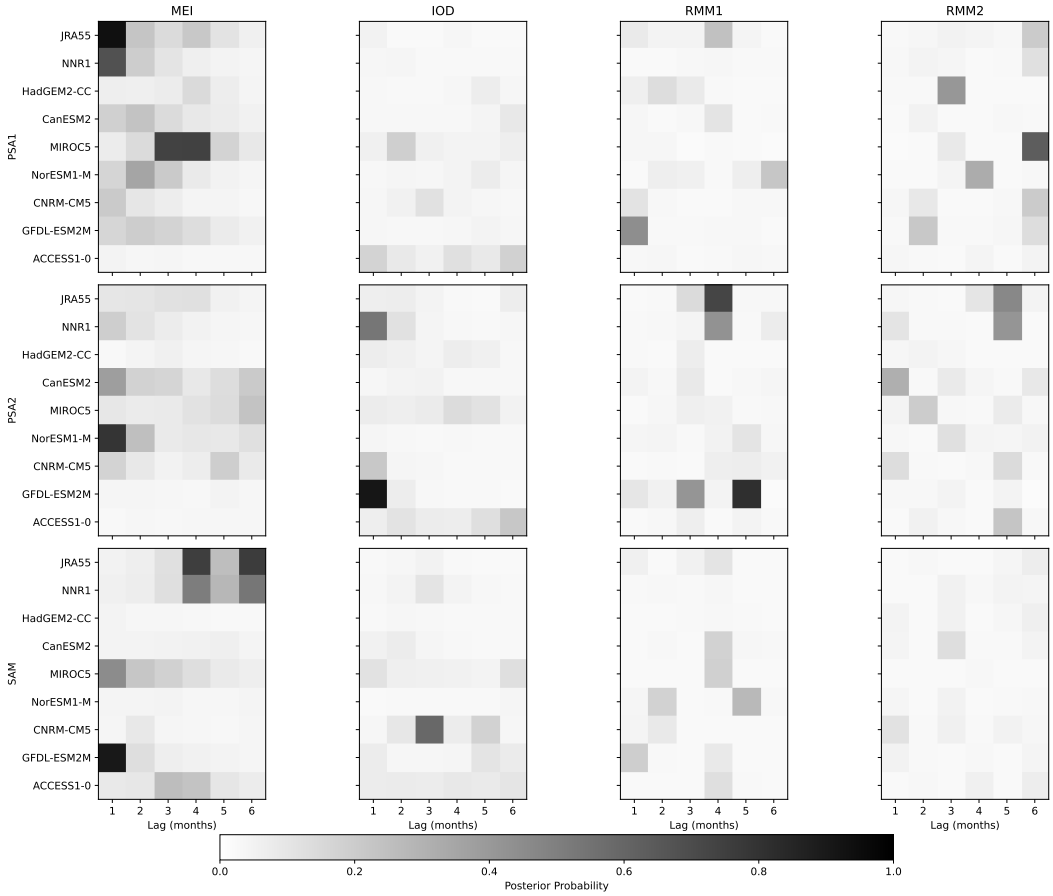

Figure 4. Heat map for the posterior weights between Southern Hemisphere (SH) indices at $t = 0$ and time-lagged tropical indices calculated over all seasons. Here the respective models shown in the rows are ordered by calculating the Wasserstein distance between models based on all posterior edge weights between the tropics and the SH simultaneously, not just the subset shown here. In this and all subsequent heatmap figures, the model ordering was determined such that the distance between adjacent models is minimized, allowing for the ordering of successive rows according to their similarity and corresponding to table 3.

In figure 4 we show the heat map for a small subset of the total posterior edge weights between the three chosen SH indices at $t = 0$ and time-lagged tropical indices calculated over all seasons. The respective models indicated in the rows are ordered based on the earthmover's distance measure summing over all tropical-SH posterior edge weights. We can see that important lagged relationships between the SH extratropics and the tropics present in the reanalyses are not captured by the CMIP5 models. In particular, both reanalyses show the known influence of ENSO on the frequency of occurrence and per-

sistence of coherent synoptic scale features in the SH mid-troposphere as represented by the PSA1 mode at lags 1 through to lag 3 (Mo, 2000; O'Kane et al., 2017) and the higher latitude westerly winds via the SAM at lags 4 through 6. While a subset of CMIP5 models have edges indicating a relationship between ENSO and the PSA1, there is very much weaker evidence for the presence of these edges than found in the reanalyses. In contrast, none of the CMIP5 models capture the time lagged influence of ENSO on the SAM apart from ACCESS1-0 where the evidence of dependence is weak and at shorter lag i.e., $t-(\tau = 3, 4)$. MIROC5 and GFDL-ESM2M have the ENSO-SAM teleconnection at $t-(\tau = 1)$ and decaying thereafter. While we can readily describe these particular cases of model biases due to their occurrence across a range of models and their ready physical interpretation, other biases represented in figure 4 are more generally model specific requiring detailed examination of the posterior weights across a number of DAGs specific to a particular model to inform where biased teleconnections may be caused by, or the cause of, related biases.

Further examples of what appear to be systematic biases across CMIP5 models occur for the tropical-NH teleconnections. As an example, all models, with the exception of CanESM2, have notable posterior edge weights for the ENSO (MEI) teleconnection to the AO extending in most cases from $t-(\tau = 1, ..., 4)$. This teleconnection is largely absent in the reanalyses over lags 1 to 3, and is only weakly supported at longer lags. The tendency for the CMIP5 models considered here to overemphasize ENSO-AO teleconnections has been previously observed in seasonal predictions using the North American Multimodel ensemble (L'Heureux et al., 2017). With some studies pointing to important regional effects on climate extremes due to specific combinations of El Niño / La Niña and the phases of the AO, and in particular over China (Chen et al., 2013), examination of this poorly understood teleconnection is becoming of increasing importance. A similar systematic bias occurs for the tropical Pacific influence on NH blocking as described by the SCAND and AR indices. Here we see for the reanalyses that this teleconnection is most strongly supported at longer lags, however, for five of the CMIP5 models an interaction with a lag of 1 month is more strongly favoured. The models and reanalyses generally are in much better agreement on the MEI-PNA teleconnection over lags up to 3 months with the exceptions of CNRM-CM5 where it is little evidenced and for NorESM1-M where the MEI influence on the PNA occurs mainly at lag $t-(\tau = 1)$.

Figure 6 shows summary DAGs for the three SH teleconnections considered namely, the PSA1 & 2 and SAM. The DAGs for both reanalyses contain the well known influence of ENSO on the PSA1 on intraseasonal timescales and on SAM at seasonal timescales. Both also show strong support for SAM autocorrelation at lag 1. Relative to JRA-55, in NNR1 there is evidence of a number of additional Granger causal relationships such as a lagged influence of the PSA2 onto the PSA1, which itself has an increased autocorrelation. NNR1 shows additional teleconnections between the NAO+ and the PSA1 at lag 4 and between the AR and SAM at lag 5. In figure 6 and the associated heat map figure 7, it is readily apparent that the CMIP5 models are diverse with only general agreement found for the SH indices autocorrelations.

Overall, and as expected, NNR1 is closest to JRA-55 with the ordering of the CMIP5 models varying according to geographic location. That said, if one considers only posterior weights > 0.5, i.e., those teleconnections between indices for which we have a high degree of confidence, then the ordering of the CMIP5 models can change markedly again. In general, it is sufficient to say that the free running CMIP5 models all exhibit systematic biases in their representation of the internal modes of variability over the recent past relative to the two reanalyses considered, both of which are quantitatively shown to be in broad agreement. In figure 8 we show the complete map of estimated posterior probabilities for the parent sets of all 13 indices, for all models and lags and where the models are ordered by their Wasserstein distances calculated over all posterior edge probabilities without thresholding. In figure 9 we show the corresponding heat map for the posterior mean $\hat{\beta}$ associated with the MAP structure for each reanalysis and CMIP5 model. As each MAP structure contains only a subset of the potential edges, these maps are necessarily sparser than those showing the posterior probabilities for each edge. Generally, the edges present in each MAP structure correspond to those with high posterior probability overall. This suggests that the edges with high posterior weight largely correspond to strong associations between parents and the respective child node, hence indicating either a strong autocorrelation or Granger causal relationship. We again emphasize that differences in a particular feature i.e., bias, requires confirmation via close examination of the representation of the processes in question in both the particular selected model and chosen reference reanalysis.

4.2 Seasonal networks for monthly indices

Harries and O'Kane (2021) previously discussed a number of seemingly spurious inter-hemispheric teleconnections in the NNR1 and JRA-55 reanalyses and the processes by which they could occur when data from all seasons is considered. Differences in the DAGs from one model to another may occur where confounding or spurious associations are generated through a failure of causal sufficiency i.e., omission of relevant variables in the fit. Seasonal variations in the background flow also profoundly influence the mechanisms that determine the variability and spatial structures of the various climate modes and their teleconnections, and in particular those at the mid- to high latitudes. This seasonal dependence should be treated appropriately by including a systematic seasonal component together with seasonal indicators as nodes within the graph (Harwood et al., 2021). Alternately one might consider time-varying network structures but this is an exceedingly challenging task and beyond the scope of the current study.

To better account for seasonality we now restrict the analysis to the boreal winter in order to capture the largest component of the interseasonal variations of the NH. We account for seasonality by restricting our analysis to data between December through to February (DJF), while still allowing for lags of up to six months such that observations entering into the fits include lagged values of the indices during the previous Autumn (SON). Previously, Harries and O'Kane (2021) showed the estimated posterior probabilities for the parent sets of the NAO^+ and NAO^- indices during DJF. In figures 10, 11 & 12 we show the DAGs associated with all of the NH indices considered namely, the NAO^+ and NAO^-, PNA, SCAND, AR and AO for both reanalyses and CMIP5 models during the boreal winter. Once again, the summarized DAGs exhibit close correspondence between JRA-55 and NNR1 apart from the parents of the NAO−. The CMIP5 models in general reproduce the AO autocorrelation and posterior edge weights from AO parents to the NAO^+ and NAO^- child nodes. However, the preferred parent sets of the other CMIP5 NH child nodes differ substantially with many of these differences arising from a general tendency to a long autocorrelation and strong lagged influence of the PNA in four of the CMIP5 models (HadGEM-CC, NorESM1, MIROC5 & CNRM-CM5) considered.

5 Summary

Recent approaches to inferring the complex interactions of the climate have largely been constraint-based (Runge, 2015; Runge et al., 2015, 2019; Spirtes et al., 2000; Nowack et al., 2020) with the advantage that it allows efficient inference for high-dimensional systems with memory. One possible drawback of the constraint-based approach is that it is hard to estimate uncertainties associated with any particular choice of structure. Uncertainty quantification on the other hand is an inherent part of Bayesian structure learning whereby the posterior distribution is learned over many possible structures rather than a single graph estimate. In the Bayesian approach, averages are taken over the set of possible models sampled from the model posterior distribution therefore allowing identification of those edges that are well supported by the data. One can then straightforwardly estimate the model parameters conditional on a given structure thereby providing a basis for comparing both the structure of the graphs and their associated parameters. Here we use MCMC algorithms to sample from the set of possible models to generate posterior probabilities and determine robust edges for which there is high confidence as a basis for climate model evaluation.

The approach we have outlined for climate model evaluation consists of three separate stages. The first stage is one of dimension reduction whereby empirical indices of the various climate modes of variability are extracted from climate model data and or observational estimates. The second stage, given a set of priors, deploys Bayesian MCMC methods to sample from the posterior distribution over possible structures followed by the third stage, where we evaluate the resulting posterior distributions for the presence of individual edges between nodes of the DAGs and the associated MAP estimates for the graph structure. Given the nodes of the resulting probabilistic graphical models are entirely specified in terms of physically observable climate modes, the graphs are at once intuitive and easily interpretable in terms of interactions between the various climate processes.

Previously Harries and O'Kane (2021) showed qualitatively that the network features derived from NNR1 and JRA-55 data with high estimated posterior probabilities are overall in good agreement. In this study we have found that this agreement is substantially better than that found when comparing any of the CMIP5 models to reanalyses, by comparing the estimated posterior probabilities for individual edges. Given the

reanalyses are constrained by observations, this is to be expected and allows for free running climate models of the historical period to be objectively compared.

Whereas differences between the models estimated from the two reanalyses are in the majority of cases limited to edges with low posterior mass, this is not the case for the CMIP5 models. For the tropical climate modes with autocorrelations that extend beyond a season (ENSO, MJO) and their parent associations, we find that the CMIP5 models are in reasonable quantitative agreement with the reanalyses. We see this general agreement start to breakdown as tropical-extratropical teleconnections are considered. The greatest diversity amongst CMIP5 models occurred when considering the mid- and high latitudes climate modes. This latitudinal dependence also corresponded to the largest divergences between the CMIP5 models and the JRA-55 reference reanalysis. While some of these feature differences in the full year fits (ALL) may arise as a result of the seasonal cycle and other common drivers, such as for periods when the midlatitude jets covary, we can test for this by observing if the apparently spurious cross-equatorial dependencies disappear when the fitting is based on seasonally dependent data.

A detailed evaluation of the mechanisms by which the identified potential biases in the respective CMIP5 model representations of the internal climate modes of variability arise is beyond the scope of the present discussion. Modeling centers conduct extensive characterization of the performance and biases of their various model configurations as part of their model development process. Having said that, the typical approach by which model biases and performance is assessed is via systematic comparison of individual metrics, usually focused on the mean climate and the large scale climate modes based on sea surface temperature differences between CMIP model and reanalysis products such as ENSO, IOD, interdecadal Pacific oscillation (IPO) and the Atlantic multidecadal oscillation (AMO) (Stoner et al., 2009; Rashid et al., 2013). Alternate common approaches include estimating the influence of temporal biases in given climate modes on specific variables e.e., precipitation and temperature (Chung et al., 2023). Due to the maturity of the data and range of available intercomparisons, we have chosen to focus on a subset of CMIP5 models however, in common with CMIP3 (Stoner et al., 2009) and CMIP5 (Rashid et al., 2013), the most recent phase 6 of CMIP (Rashid et al., 2022) reveals that, whereas the spatial structures of the large scale oceanic climate modes (ENSO, IOD, IPO and AMO) compare favourably with the structures of their observed counterparts, there remain major and systematic differences in the simulated temporal variability. As we have

seen in the results presented here, these biases in the temporal variability of the tropical modes and their teleconnections to the midlatitude atmospheric modes are a major source of model error.

Building on the previous initial application of DBNs to reanalysis data by Harries and O'Kane (2021), we have shown additional evidence for consistency across the JRA-55 and NNR1 reanalyses confirming that a score-based approach recovers the expected teleconnections between the climate modes through estimation of the posterior distribution over models and features. The DBN models obtained for the NNR1 and JRA-55 reanalyses have been used as a set of ground truth results against which free-running CMIP5 models have been compared over the historical period. The results presented here further indicate that systematic biases exist across a broad range of climate model configurations largely in the temporal variability of the major atmospheric modes of variability and their teleconnections. These biases tend to be somewhat mitigated where there are well defined teleconnections from the tropics to the extratropics. The tropical modes, defined in terms of indices based on sea surface temperature with the longest autocorrelation are in the best agreement to the reanalyses.

It is important to distinguish between systematic model error and temporally dependent differences due to the phases of the large scale background state. Shown in figure 13 are the dependencies (parents) for the tropical indices (children) present in the JRA-55 data over the period 1958-10-01 to 1998-12-31 (including all seasons) encompassing two phases of the inter-decadal Pacific oscillation IPO: (Power et al., 1999). Specifically we show dependencies for the corresponding negative (-IPO: 1958-10-01 to 1976-12-31) and positive (+IPO: 1977-01-01 to 1998-12-31) IPO phases. Here we see very similar dependencies between those calculated over the entire period and those over the +IPO phase. For the earlier period corresponding to the -IPO phase, we observed increased auto-correlations for ENSO (MEI) but significantly weaker auto-correlations for the MJO, and in particular for the RMM2 component of the real-time multivariate Madden-Julian oscillation index, relative to the full period. In addition, a significant teleconnection emerges between the SAM and tropics in this latter period consistent with observed increased strength of the mid- to high latitude westerly winds in the SH. Similarly, we observe the emergence of a stronger PSA1 teleconnection to the tropics. Overall, the +IPO phase DAG more closely resembles that of the full period. On the basis of this and associated investigations, we posit that the largest differences between the dependency structures

of the reanalyses and climate models are in fact due to model errors and biases. There is however undoubtedly an additional component due to temporal variations due to systematic changes in the background state resulting in the aforementioned structural differences between the -IPO and +IPO DAGs. Whereas, the free-running models are unconstrained and not able to reproduce the observed temporal changes due to the phase relationships within a given regime, the models will diverge from the true trajectory.

Here we have considered the restricted case of homogeneous models derived from monthly mean data and in doing so ignore secular trends in the spatio-temporal evolution of the modes. O'Kane et al. (2016) have shown that the SH midlatitude atmospheric modes in the SH have undergone systematic changes in their structure, frequency of occurrence and persistence over the recent historical record. In order to understand the impact of structural changes in the persistent states of the observed or future anthropogenically forced troposphere, on teleconnections to the wider climate system requires models that are inherently constructed to handle nonstationarity and with regime identification. Saggioro et al. (2020) proposed a potentially useful framework for embedding homogeneous network models within a variational approach to regime identification. In the future we intend to extend homogeneous network models through the inclusion of regime identification to better understand the response of the simulated climate system to changes in radiative forcing, be they slow systematic variations or abrupt changes in the underlying network structure indicative of regime transitions. Furthermore, the considered random variables, be they climate indices or any other time series, should be chosen with care as the "important/main" variables for the phenomena under investigation as with any causal algorithm it is possible that the inferred structural causal model would change when adding more indices to the analysis (Baldovin et al., 2020).

Our purpose here is not simply to demonstrate how Bayesian structure learning can be of utility as a tool for process-based model evaluation, but that it affords a unique approach whereby one can assess biases in the temporal behaviour of individual climate modes and identify and assess the teleconnections between those modes. As the analytic posterior distributions can be factorized, the associated DAGs can be fitted simultaneously and computationally efficiently across a large representative sample of climate model simulations. In summary, Bayesian structural causal models naturally afford uncertainty estimation in order to ascertain the robustness of differences across models and observations and hence identify genuine model biases.

Acknowledgments

The authors very much appreciate the efforts of two anonymous reviewers and the handling editor for their valuable comments and suggestions.

The HadISST SST dataset is provided by the UK Met Office Hadley Centre as described in Rayner et al. (2003), and may be accessed at `https://www.metoffice.gov.uk/hadobs/hadisst/` (last access: 29 April 2019). The NCEP/NCAR reanalysis output used is provided by the NOAA/OAR/ESRL PSL, Boulder, Colorado, USA, described in Kalnay et al. (1996), and may be accessed at `https://psl.noaa.gov/data/reanalysis/reanalysis.shtml` (last access: 10 May 2019). The JRA-55 reanalysis output used is made available through the JRA-55 project and may be accessed following the procedures and access conditions described in Kobayashi et al. (2015) and at `https://jra.kishou.go.jp/JRA-55/index_en.html` (last access: 12 April 2019). The CMIP5 data used in this study can be accessed via Program for Climate Model Diagnosis & Iintercomparison (PCMDI) at https://pcmdi.llnl.gov/mips/cmip5/.

Regridding of the reanalysis fields was performed using the Climate Data Operators software suite (Schulzweida, 2019), while the analysis code was implemented using the Python libraries NumPy (Oliphant, 2006; Van Der Walt et al., 2011), SciPy (Virtanen et al., 2020), pandas (Wes McKinney, 2010), scikit-learn (Pedregosa et al., 2011), and xarray (Hoyer & Hamman, 2017). Plots were generated using the Python package Matplotlib (Hunter, 2007). All source code for the algorithms used to perform the analyses, including those needed to generate the climate indices, and results presented in this study may be found at `https://doi.org/10.5281/zenodo.4331149`.

DH is supported by the South Australian Health and Medical Research Institute. TJO and MAC are supported by CSIRO.

Appendix A Sampling algorithms

As noted in the main text, inference for DBNs with a priori unknown structure may be separated into two phases, based on decomposing the joint posterior distribution for the model structure G and parameters θ. To account for model uncertainty, we aim to generate a sample of structures G from the posterior distribution

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)}, \tag{A1}$$

given data D. The marginal likelihood, Eq. (4), is obtained by integrating out the model parameters. For the DBN models used in this study, the likelihood under a given structure G, corresponding to a particular choice of the parent sets

$$\mathrm{pa}_G(Y_t^i) = \{Y_{t-\tau}^j | G \text{ contains an edge from } Y_{t-\tau}^j \text{ to } Y_t^i\}, \tag{A2}$$

can be written

$$P(D|G,\theta) = \prod_{t=1}^{T}\prod_{i=1}^{n} P(Y_t^i|\mathrm{pa}_G(Y_t^i),\theta_i); \tag{A3}$$

we assume that a sufficiently large set of pre-sample values have been held out to condition on. Assuming that the priors appearing in Eq. (4) for the model parameters satisfy the properties of parameter independence (Heckerman et al., 1995),

$$P(\theta|G) = \prod_{i=1}^{n} P(\theta_i|G), \tag{A4}$$

and modularity (that is, for any two graphs G and G', if Y_t^i has the same parent set in G and G', then the priors for the parameters θ_i characterizing the conditional PDF of Y_t^i satisfy $P(\theta_i|G) = P(\theta_i|G')$), the marginal likelihood may be written as the product of local marginal likelihoods $\Psi_i(D,G)$ (Grzegorczyk & Husmeier, 2011):

$$P(D|G) = \prod_{i=1}^{n}\int d\theta_i \prod_{t=1}^{T} P(Y_t^i|\mathrm{pa}_G(Y_t^i),\theta_i)P(\theta_i|G) \equiv \prod_{i=1}^{n}\Psi_i(D;G). \tag{A5}$$

For structurally modular priors of the form

$$P(G) = \prod_{i=1}^{n} P(\mathrm{pa}_G(Y_t^i)) \tag{A6}$$

the posterior over graphs also factorizes,

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} = \frac{1}{P(D)}\prod_{i=1}^{n}\Psi_i(D;G)P(\mathrm{pa}_G(Y_t^i)), \tag{A7}$$

so that each factor can be computed independently, up to an overall normalization.

For general choices of the conditional densities $P(Y_t^i|\mathrm{pa}_G(Y_t^i),\theta_i)$, it is not possible to analytically marginalize out the model parameters (i.e., evaluate $\Psi_i(D;G)$ in closed-form), as would be required to sample from the marginal posterior distribution $P(G|D)$ directly. Instead, it is necessary to construct a MCMC sampler that samples from the joint posterior $P(\theta,G|D)$ using, e.g., reversible jump Markov Chain Monte Carlo (RJMCMC) (Green, 1995) or related methods (Carlin & Chib, 1995; Godsill, 2001).

Methods for sampling from the space of possible structures may be neatly formulated in terms of the composite parameter space formulation of Godsill (2001). From the

collection of parameters associated with all allowable models, $\boldsymbol{\theta}$, any given model G will depend only on some subset $\boldsymbol{\theta}_{\mathcal{I}(G)}$. Sampling from the joint posterior distribution for model structures and parameters can be performed using a Metropolis-Hastings type scheme on a composite parameter space. Briefly, at each iteration either: 1) with probability $j_\theta(G, \boldsymbol{\theta}_{\mathcal{I}(G)})$, a new set of parameters associated with the graph structure G is proposed, or 2) an update to the current structure is proposed with probability $1-j_\theta(G, \boldsymbol{\theta}_{\mathcal{I}(G)})$. Where the model structure is left unchanged, $G' = G$, and a new set of parameter values $\boldsymbol{\theta}'_{\mathcal{I}(G)}$ is drawn from a proposal density $q_\theta(\boldsymbol{\theta}'_{\mathcal{I}(G)}; \boldsymbol{\theta}_{\mathcal{I}(G)})$. The new state, consisting of the structure G and proposed new parameter values $\boldsymbol{\theta}'_{\mathcal{I}(G)}$, is accepted with probability

$$\alpha = \min\left\{1, \frac{j_\theta(G, \boldsymbol{\theta}'_{\mathcal{I}(G)})}{j_\theta(G, \boldsymbol{\theta}_{\mathcal{I}(G)})} \frac{q_\theta(\boldsymbol{\theta}_{\mathcal{I}(G)}; \boldsymbol{\theta}'_{\mathcal{I}(G)})}{q_\theta(\boldsymbol{\theta}'_{\mathcal{I}(G)}; \boldsymbol{\theta}_{\mathcal{I}(G)})} \frac{P(G, \boldsymbol{\theta}'_{\mathcal{I}(G)}|D)}{P(G, \boldsymbol{\theta}_{\mathcal{I}(G)}|D)}\right\}. \tag{A8}$$

Alternatively, if an update to the current structure is to be made, a new structure G' is drawn according to a proposal distribution $q_G(G'; G)$, and any new parameters required to fully specify G', $\boldsymbol{\theta}'_{\mathcal{I}(G')\setminus\mathcal{I}(G)}$, are drawn from a proposal density $\tilde{q}_\theta(\boldsymbol{\theta}'_{\mathcal{I}(G')\setminus\mathcal{I}(G)})$. All other parameters are retained at their previous values. The new state is accepted with probability

$$\alpha = \min\left\{1, \frac{j_G(G', \boldsymbol{\theta}'_{\mathcal{I}(G')})}{j_G(G, \boldsymbol{\theta}_{\mathcal{I}(G)})} \frac{q_G(G; G')}{q_G(G'; G)} \frac{\tilde{q}_\theta(\boldsymbol{\theta}_{\mathcal{I}(G)\setminus\mathcal{I}(G')})}{\tilde{q}_\theta(\boldsymbol{\theta}'_{\mathcal{I}(G')\setminus\mathcal{I}(G)})} \frac{P(G', \boldsymbol{\theta}'_{\mathcal{I}(G')}|D)}{P(G, \boldsymbol{\theta}_{\mathcal{I}(G)}|D)}\right\}. \tag{A9}$$

For models where the conditional posterior distribution for all parameters admits analytic evaluation the above scheme reduces to the MC^3 scheme of Madigan et al. (1995) (see Algorithm 2 of Harries and O'Kane (2021)). The acceptance ratio for a structure drawn according to $q_G(G'; G)$ is in this case given by Eq. (5). We use this sampling algorithm for all of the results presented in this study. The required closed-form expressions for the prior and posterior densities for the parameters of the node conditional distributions, and the resulting marginal likelihoods or local scores, for the models used are detailed in Appendix B of Harries and O'Kane (2021).

References

Arnold, A., Liu, Y., & Abe, N. (2007). Temporal Causal Modeling with Graphical Granger Methods. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (p. 66–75). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/1281192.1281203

Bach, E., Motesharrei, S., Kalnay, E., & Ruiz-Barradas, A. (2019). Local atmosphere-ocean predictability: Dynamical origins, lead times, and seasonality. *J. Climate*, *32*, 7507 - 7519. doi: 10.1175/JCLI-D-18-0817.1

Baldovin, M., Cecconi, F., & Vulpiani, A. (2020). Understanding causation via correlations and linear response theory. *PHYSICAL REVIEW RESEARCH*, *2*, 043436. doi: PhysRevResearch.2.043436

Barnston, A. G., & Livezey, R. E. (1987). Classification, Seasonality and Persistence of Low-Frequency Atmospheric Circulation Patterns. *Monthly Weather Review*, *115*(6), 1083-1126. doi: 10.1175/1520-0493(1987)115⟨1083:CSAPOL⟩2.0.CO;2

Bello, G. A., Angus, M., Pedemane, N., Harlalka, J. K., Semazzi, F. H. M., Kumar, V., & Samatova, N. F. (2015). Response-Guided Community Detection: Application to Climate Index Discovery. In *Machine Learning and Knowledge Discovery in Databases* (pp. 736–751). Cham: Springer International Publishing. doi: 10.1007/978-3-319-23525-7_45

Bentsen, M., Bethke, I., Debernard, J., Iversen, T., Kirkeväg, A., Seland, O., ... Kristjánsson, J. E. (2013). The Norwegian Earth System Model, NorESM1-M Part1:Description and basic evaluation of the physical climate. *Geosci.Model Dev.*, *6*, 687–720. doi: 10.5194/gmd-6-687-2013

Bi, D., Dix, M., Marsland, S. J., O'Farrell, S., Rashid, H. A., Uotila, P., ... Puri, K. (2013). The ACCESS coupled model: description, control climate and evaluation. *Aust. Meteorol. Ocean. J.*, *63*, 41–64.

Bjerknes, J. (1969). Atmospheric Teleconnections from the Equatorial Pacific. *Monthly Weather Review*, *97*(3), 163-172. doi: 10.1175/1520-0493(1969)097⟨0163:ATFTEP⟩2.3.CO;2

Bromwich, D. H., & Fogt, R. L. (2004). Strong Trends in the Skill of the ERA-40 and NCEP–NCAR Reanalyses in the High and Midlatitudes of the Southern Hemisphere, 1958–2001. *Journal of Climate*, *17*(23), 4603-4619. doi: 10.1175/3241.1

Bromwich, D. H., Fogt, R. L., Hodges, K. I., & Walsh, J. E. (2007). A tropospheric assessment of the ERA-40, NCEP, and JRA-25 global reanalyses in the polar regions. *Journal of Geophysical Research: Atmospheres*, *112*(D10). doi: 10.1029/2006JD007859

Brooks, S., Giudici, P., & Philippe, A. (2003). Nonparametric Convergence Assessment for MCMC Model Selection. *Journal of Computational and Graphical Statistics*, *12*(1), 1-22. doi: 10.1198/1061860031347

Burnham, K. P., & Anderson, D. R. (2002). *Model Selection and Multi-Model Inference* (2nd edition ed.). Springer.

Carlin, B. P., & Chib, S. (1995). Bayesian Model Choice Via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(3), 473-484. doi: 10.1111/j.2517-6161.1995.tb02042.x

Chen, W., Lan, X., & Ma, Y. (2013). The combined effects of the ENSO and the Arctic Oscillation on the winter climate anomalies in East Asia. *Chinese Science Bulletin*, *58*(12), 1355–1362. doi: 10.1007/s11434-012-5654-5

Chung, C., Boschat, G., Taschetto, A., Narsey, S., McGregor, S., Santoso, A., & Delage, F. (2023). Evaluation of seasonal teleconnections to remote drivers of Australian rainfall in CMIP5 and CMIP6 models. *Journal of Southern Hemisphere Earth Systems Science*. doi: 10.1071/ES23002

Cooper, G. F., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning*, *9*(4), 309–347.

Donges, J. F., Zou, Y., Marwan, N., & Kurths, J. (2009a). The backbone of the climate network. *EPL (Europhysics Letters)*, *87*(4), 48007. doi: 10.1209/0295-5075/87/48007

Donges, J. F., Zou, Y., Marwan, N., & Kurths, J. (2009b). Complex networks in climate dynamics. *The European Physical Journal Special Topics*, *174*(1), 157–179. doi: 10.1140/epjst/e2009-01098-2

Draper, D. (1995). Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 45-70. doi: 10.1111/j.2517-6161.1995.tb02015.x

Dunne, J., John, J., Shevliakova, E., Stouffer, R., Krasting, J., Malyshev, S., ... Zadeh, N. (2013). GFDL's ESM2 global coupled climate–carbon earth system models. Part ii: Carbon system formulation and baseline simulation characteristics. *J. Climate*, *26*, 2247–2267. doi: 10.1175/JCLI-D-12-00150.1

Eichler, M. (2012). Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, *153*(1), 233–268. doi: 10.1007/s00440-011-0345-8

Falasca, F., Bracco, A., Nenes, A., & Fountalis, I. (2019). Dimensionality Reduc-

tion and Network Inference for Climate Data Using δ-MAPS: Application to the CESM Large Ensemble Sea Surface Temperature. *Journal of Advances in Modeling Earth Systems*, *11*(6), 1479-1515. doi: 10.1029/2019MS001654

Friedman, N., & Koller, D. (2003). Being Bayesian About Network Structure. A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning*, *50*(1), 95–125. doi: 10.1023/A:1020249912095

Friedman, N., Murphy, K., & Russell, S. (1998). Learning the Structure of Dynamic Probabilistic Networks. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence* (p. 139–147). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Geiger, D., & Heckerman, D. (1994). Learning gaussian networks. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence* (pp. 235–243).

Geng, T., Jia, F., Cai, W., Wu, L., Gan, B., Jing, Z., ... McPhaden, M. J. (2023). Increased occurrences of consecutive la niña events under global warming. *Nature*, *619*, 774–781. doi: 10.1038/s41586-023-06236-9

Godsill, S. J. (2001). On the Relationship Between Markov chain Monte Carlo Methods for Model Uncertainty. *Journal of Computational and Graphical Statistics*, *10*(2), 230-248. doi: 10.1198/10618600152627924

Greatbatch, R. J., & Rong, P.-p. (2006). Discrepancies between Different Northern Hemisphere Summer Atmospheric Data Products. *Journal of Climate*, *19*(7), 1261-1273. doi: 10.1175/JCLI3643.1

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711-732. doi: 10.1093/biomet/82.4.711

Grzegorczyk, M., & Husmeier, D. (2011). Non-homogeneous dynamic Bayesian networks for continuous data. *Machine Learning*, *83*(3), 355–419. doi: 10.1007/s10994-010-5230-7

Hall, P. (1987). On kullback-leibler loss and density estimation. *Ann. Statist.*, *15*, 1491–1519. doi: 10.1214/aos/1176350606

Harada, Y., Kamahori, H., Kobayashi, C., Endo, H., Kobayashi, S., Ota, Y., ... Takahashi, K. (2016). The JRA-55 Reanalysis: Representation of Atmospheric Circulation and Climate Variability. *Journal of the Meteorological Society of*

Japan. Ser. II, *94*(3), 269-302. doi: 10.2151/jmsj.2016-015

Harries, D., & O'Kane, T. J. (2021). Dynamic bayesian networks for evaluation of granger causal relationships in climate reanalyses. *Journal of Advances in Modeling Earth Systems*, *13*, e2020MS002442. doi: 10.1029/2020MS002442

Harwood, N., Hall, R., Di Capua, G., Russell, A., & Tucker, A. (2021). Using Bayesian Networks to Investigate the Influence of Subseasonal Arctic Variability on Midlatitude North Atlantic Circulation. *Journal of Climate*, *34*(6), 2319 - 2335. doi: 10.1175/JCLI-D-20-0369.1

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine learning*, *20*(3), 197–243.

Hertzog, A., Basdevant, C., & Vial, F. (2006). An Assessment of ECMWF and NCEP–NCAR Reanalyses in the Southern Hemisphere at the End of the Pre-satellite Era: Results from the EOLE Experiment (1971–72). *Monthly Weather Review*, *134*(11), 3367-3383. doi: 10.1175/MWR3256.1

Hines, K. M., Bromwich, D. H., & Marshall, G. J. (2000). Artificial Surface Pressure Trends in the NCEP–NCAR Reanalysis over the Southern Ocean and Antarctica. *Journal of Climate*, *13*(22), 3940-3952. doi: 10.1175/1520-0442(2000)013⟨3940:ASPTIT⟩2.0.CO;2

Horel, J. D., & Wallace, J. M. (1981). Planetary-Scale Atmospheric Phenomena Associated with the Southern Oscillation. *Monthly Weather Review*, *109*(4), 813-829. doi: 10.1175/1520-0493(1981)109⟨0813:PSAPAW⟩2.0.CO;2

Hoyer, S., & Hamman, J. (2017). xarray: N-D labeled arrays and datasets in Python. *Journal of Open Research Software*, *5*(1). doi: 10.5334/jors.148

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. doi: 10.1109/MCSE.2007.55

IPCC. (2013). *The Physical Science Basis. Contribution ofWorking Group I to the Fifth Assessment Report of the IntergovernmentalPanel on Climate Change* (Tech. Rep.). Cambridge University Press, UnitedKingdom and New York.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187–200. doi: 10.1007/BF02289233

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., ... Joseph, D. (1996). The NCEP/NCAR 40-Year Reanalysis Project.

Bulletin of the American Meteorological Society, *77*(3), 437-472. doi: 10.1175/1520-0477(1996)077⟨0437:TNYRP⟩2.0.CO;2

Kistler, R., Kalnay, E., Collins, W., Saha, S., White, G., Woollen, J., ... Fiorino, M. (2001). The NCEP–NCAR 50-Year Reanalysis: Monthly Means CD-ROM and Documentation. *Bulletin of the American Meteorological Society*, *82*(2), 247-268. doi: 10.1175/1520-0477(2001)082⟨0247:TNNYRM⟩2.3.CO;2

Kitsios, V., O'Kane, T. J., & Zagar, N. (2019). A Reduced-Order Representation of the Madden–Julian Oscillation Based on Reanalyzed Normal Mode Coherences. *Journal of the Atmospheric Sciences*, *76*, 2463–2480. doi: 10.1175/JAS-D-18-0197.1

Kjærulff, U. (1995). dhugin: A computational system for dynamic time-sliced bayesian networks. *International journal of forecasting*, *11*(1), 89–111.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., ... Takahashi, K. (2015). The JRA-55 Reanalysis: General Specifications and Basic Characteristics. *Journal of the Meteorological Society of Japan. Ser. II*, *93*(1), 5-48. doi: 10.2151/jmsj.2015-001

Kretschmer, M., Runge, J., & Coumou, D. (2017). Early prediction of extreme stratospheric polar vortex states based on causal precursors. *Geophysical Research Letters*, *44*(16), 8592-8600. doi: 10.1002/2017GL074696

Lau, K.-M., Sheu, P.-J., & Kang, I.-S. (1994). Multiscale Low-Frequency Circulation Modes in the Global Atmosphere. *Journal of the Atmospheric Sciences*, *51*(9), 1169-1193. doi: 10.1175/1520-0469(1994)051⟨1169:MLFCMI⟩2.0.CO;2

Lèbre, S. (2009). Inferring Dynamic Genetic Networks with Low Order Independencies. *Statistical Applications in Genetics and Molecular Biology*, *8*, 1 – 38. doi: 10.2202/1544-6115.1294

Lèbre, S., Becq, J., Devaux, F., Stumpf, M. P., & Lelandais, G. (2010). Statistical inference of the time-varying structure of gene-regulation networks. *BMC Systems Biology*, *4*(130), 1. doi: 10.1186/1752-0509-4-130

Lindsay, R., Wensnahan, M., Schweiger, A., & Zhang, J. (2014). Evaluation of Seven Different Atmospheric Reanalysis Products in the Arctic*. *Journal of Climate*, *27*(7), 2588-2606. doi: 10.1175/JCLI-D-13-00014.1

Lorenz, E. N. (1956). *Empirical Orthogonal Functions and Statistical Weather Prediction* (Tech. Rep.). Cambridge: Massachusetts Institute of Technology.

L'Heureux, M. L., Tippett, M. K., Kumar, A., Butler, A. H., Ciasto, L. M., Ding, Q., ... Johnson, N. C. (2017). Strong relations between ENSO and the Arctic Oscillation in the NorthAmerican Multimodel Ensemble. *Geophysical Research Letters*, *44*, 11,654–11,662. doi: 10.1002/2017GL074854

Madden, R. A., & Julian, P. R. (1971). Detection of a 40–50 day oscillation in the zonal wind in the tropical pacific. *J. Atmos. Sci.*, *28*, 702–708. doi: 10.1175/1520-0469(1971)028,0702:DOADOI.2.0.CO;2

Madigan, D., & Raftery, A. E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, *89*(428), 1535-1546. doi: 10.1080/01621459.1994.10476894

Madigan, D., York, J., & Allard, D. (1995). Bayesian Graphical Models for Discrete Data. *International Statistical Review / Revue Internationale de Statistique*, *63*(2), 215–232. doi: 10.2307/1403615

Marshall, G. J. (2002). Trends in Antarctic Geopotential Height and Temperature: A Comparison between Radiosonde and NCEP–NCAR Reanalysis Data. *Journal of Climate*, *15*(6), 659-674. doi: 10.1175/1520-0442(2002)015⟨0659:TIAGHA⟩2.0.CO;2

Marshall, G. J., & Harangozo, S. A. (2000). An appraisal of NCEP/NCAR reanalysis MSLP data viability for climate studies in the South Pacific. *Geophysical Research Letters*, *27*(19), 3057-3060. doi: 10.1029/2000GL011363

Martin, G. M., Bellouin, N., Collins, W. J., Culverwell, I. D., Halloran, P. R., Hardiman, S. C., ... Wiltshire, A. (2011). The hadgem2 family of met office unified model climate configurations. *Geosci. Model Dev.*, *4*, 723–757. doi: 10.5194/gmd-4-723-2011

McGraw, M. C., & Barnes, E. A. (2018). Memory Matters: A Case for Granger Causality in Climate Variability Studies. *Journal of Climate*, *31*(8), 3289-3300. doi: 10.1175/JCLI-D-17-0334.1

Mo, K. C. (2000). Relationships between low-frequency variability in the Southern Hemisphere and sea surface temperature anomalies. *J. Climate*, *13*, 3599–3610.

Mo, K. C., & Ghil, M. (1987). Statistics and Dynamics of Persistent Anomalies. *Journal of the Atmospheric Sciences*, *44*(5), 877-902. doi: 10.1175/

1520-0469(1987)044⟨0877:SADOPA⟩2.0.CO;2

Murphy, K., & Mian, S. (1999). *Modelling gene expression data using dynamic Bayesian networks* (Tech. Rep.). Berkely, CA: Computer Science Division, University of California.

Murphy, K. P., & Russell, S. (2002). *Dynamic Bayesian networks: representation, inference and learning*. University of California, Berkeley Dissertation.

Nowack, P., Runge, J., Eyring, V., & Haigh, J. D. (2020). Causal networks for climate model evaluation and constrained projections. *Nature Communications*, *11*(1), 1415. doi: 10.1038/s41467-020-15195-y

O'Kane, T. J., Monselesan, D. P., & Risbey, J. S. (2017). A multiscale reexamination of the Pacific South American pattern. *Mon. Wea. Rev.*, *145*(1), 379–402. doi: 10.1175/MWR-D-16-0291.1

O'Kane, T. J., Risbey, J. S., & Monselesan, D. P. (2017). A multiscale reexamination of the pacific–south american pattern. *Mon. Wea. Rev.*, *145*, 379–402. doi: 10.1175/MWR-D-16-0291.1

O'Kane, T. J., Risbey, J. S., Monselesan, D. P., Horenko, I., & Franzke, C. L. E. (2016). On the dynamics of persistent states and their secular trends in the waveguides of the Southern Hemisphere troposphere. *Climate Dynamics*, *46*(11-12), 3567–3597. doi: 10.1007/s00382-015-2786-8

Oliphant, T. E. (2006). *A guide to numpy* (Vol. 1). Trelgol Publishing USA.

Onogi, K., Tsutsui, J., Koide, H., Sakamoto, M., Kobayashi, S., Hatsushika, H., ... Taira, R. (2007). The JRA-25 reanalysis. *Journal of the Meteorological Society of Japan*, *85*(3), 369-432. doi: 10.2151/jmsj.85.369

O'Kane, T. J., Sandery, P. A., Kitsios, V., Sakov, P., Chamberlain, M. A., Squire, D. T., ... Matear, R. J. (2021). Cafe60v1: A 60-year large ensemble climate reanalysis. Part II: Evaluation. *J. Climate*, *34*, 1571–1594.

O'Kane, T. J., Squire, D. T., Sandery, P. A., Kitsios, V., Matear, R. J., Moore, T. S., ... Watterson, I. G. (2020). Enhanced ENSO prediction via augmentation of multimodel ensembles with initial thermocline perturbations. *J. Climate*, *33*, 2281–2293. doi: 10.1175/JCLI-D-19-0444.1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Power, S., Casey, T., Folland, C., Colman, A., & Mehta, V. (1999). Inter-decadal modulation of the impact of enso on australia. *Climate Dyn.*, *15*, 319–324. doi: 10.1007/s003820050284

Punskaya, E., Andrieu, C., Doucet, A., & Fitzgerald, W. J. (2002). Bayesian curve fitting using MCMC with applications to signal segmentation. *IEEE Transactions on Signal Processing*, *50*(3), 747-758. doi: 10.1109/78.984776

Rashid, H., Sullivan, A., Dix, M., Bi, D., Mackallah, C., Ziehn, T., ... Marsland, S. (2022). Evaluation of climate variability and change in ACCESS historical simulations for CMIP6. *Australian Meteorological and Oceanographic Journal*, *72*(2), 73–92. doi: 10.1071/ES21028

Rashid, H., Sullivan, A., Hirst, A., Bi, D., Zhou, X., & Marsland, S. (2013). Evaluation of El Niño–Southern Oscillation in the ACCESS coupled model simulations for CMIP5. *Australian Meteorological and Oceanographic Journal*, *63*, 161–180. doi: 10.1071/ES13010

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., ... Kaplan, A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres*, *108*(D14). doi: 10.1029/2002JD002670

Rogers, J. C., & van Loon, H. (1982). Spatial Variability of Sea Level Pressure and 500 mb Height Anomalies over the Southern Hemisphere. *Monthly Weather Review*, *110*(10), 1375-1392. doi: 10.1175/1520-0493(1982)110⟨1375:SVOSLP⟩2.0.CO;2

Runge, J. (2015). Quantifying information transfer and mediation along causal pathways in complex systems. *Phys. Rev. E*, *92*, 062829. doi: 10.1103/PhysRevE.92.062829

Runge, J. (2018). Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *28*(7), 075310. doi: 10.1063/1.5025050

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*(11). doi: 10.1126/sciadv.aau4996

Runge, J., Petoukhov, V., Donges, J. F., Hlinka, J., Jajcay, N., Vejmelka, M., ...

Kurths, J. (2015). Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature communications*, *6*(1), 1–10. doi: 10.1038/ncomms9502

Saggioro, E., de Wiljes, J., Kretschmer, M., & Runge, J. (2020). Reconstructing regime-dependent causal relationships from observational time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *30*(11), 113115. doi: 10.1063/5.0020538

Saji, N. H., Goswami, B. N., Vinayachandran, P. N., & Yamagata, T. (1999). A dipole mode in the tropical Indian Ocean. *Nature*, *401*(6751), 360–363. doi: 10.1038/43854

Schulzweida, U. (2019, October). *CDO User Guide.* Retrieved from `https://doi.org/10.5281/zenodo.3539275` doi: 10.5281/zenodo.3539275

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search. adaptive computation and machine learning* (2nd edition ed.). MIT Pres.

Steinhaeuser, K., Chawla, N. V., & Ganguly, A. R. (2011). Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *4*(5), 497-511. doi: 10.1002/sam.10100

Steinhaeuser, K., Ganguly, A. R., & Chawla, N. V. (2012). Multivariate and multiscale dependence in the global climate system revealed through complex networks. *Climate Dynamics*, *39*(3), 889–895. doi: 10.1007/s00382-011-1135-9

Stoner, A. M., Hayhoe, K., & Wuebbles, D. J. (2009). Assessing general circulation model simulations of atmospheric teleconnection patterns. *J. Climate*, *22*, 4348–4372. doi: 10.1175/2009JCLI2577.1

Straus, D. M., Molteni, F., & Corti, S. (2017). Atmospheric regimes: The link between weather and the large scale circulation. In C. L. E. Franzke & T. J. O'Kane (Eds.), *Nonlinear and stochastic climate dynamics* (pp. 105–135). Cambridge University Press, Cambridge.

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of cmip5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*(4), 485-498. doi: 10.1175/BAMS-D-11-00094.1

Thompson, D. W. J., & Wallace, J. M. (1998). The Arctic oscillation signature in the wintertime geopotential height and temperature fields. *Geophysical Re-*

search Letters, *25*(9), 1297-1300. doi: 10.1029/98GL00950

Thompson, D. W. J., & Wallace, J. M. (2000). Annular Modes in the Extratropical Circulation. Part I: Month-to-Month Variability. *Journal of Climate*, *13*(5), 1000-1016. doi: 10.1175/1520-0442(2000)013⟨1000:AMITEC⟩2.0.CO;2

Tsonis, A. A., & Roebber, P. (2004). The architecture of the climate network. *Physica A: Statistical Mechanics and its Applications*, *333*, 497 - 504. doi: 10.1016/j.physa.2003.10.045

Tsonis, A. A., & Swanson, K. L. (2008). Topology and Predictability of El Niño and La Niña Networks. *Phys. Rev. Lett.*, *100*, 228502. doi: 10.1103/PhysRevLett.100.228502

Tsonis, A. A., Swanson, K. L., & Wang, G. (2008). On the Role of Atmospheric Teleconnections in Climate. *Journal of Climate*, *21*(12), 2990-3001. doi: 10.1175/2007JCLI1907.1

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, *13*(2), 22.

van Loon, H., & Rogers, J. C. (1978). The Seesaw in Winter Temperatures between Greenland and Northern Europe. Part I: General Description. *Monthly Weather Review*, *106*(3), 296-310. doi: 10.1175/1520-0493(1978)106⟨0296:TSIWTB⟩2.0.CO;2

Vázquez-Patiño, A., Campozano, L., Mendoza, D., & Samaniego, E. (2020). A causal flow approach for the evaluation of global climate models. *International Journal of Climatology*, *40*(10), 4497-4517. doi: 10.1002/joc.6470

Villani, C. (2009). The Wasserstein distances. In *Optimal transport: Old and new* (pp. 93–111). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from `https://doi.org/10.1007/978-3-540-71050-9_6` doi: 10.1007/978-3-540-71050-9_6

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. doi: 10.1038/s41592-019-0686-2

Voldoire, A., Sanchez-Gomez, E., y Mélia, D. S., Decharme, B., Cassou, C., Sénési, S., ... Chauvin, F. (2013). The CNRM-CM5.1 global climate model: de-

scription and basic evaluation. *Climate Dynamics*, *40*, 2091–2121. doi: 10.1007/s00382-011-1259-y

Walker, G. T. (1923). Correlation in Seasonal Variations of Weather, VIII, a preliminary study of world weather. *Memoirs of the India Meteorological Department*, *24*, 75–131.

Walker, G. T. (1924). Correlations in Seasonal Variations of Weather. I. A further study of world weather. *Memoirs of the India Meteorological Department*, *24*, 275–332.

Wallace, J. M., & Gutzler, D. S. (1981). Teleconnections in the Geopotential Height Field during the Northern Hemisphere Winter. *Monthly Weather Review*, *109*(4), 784-812. doi: 10.1175/1520-0493(1981)109⟨0784:TITGHF⟩2.0.CO;2

Watanabe, M., Suzuki, T., O'ishi, R., Komuro, Y., Watanabe, S., Emori, S., ... Kimoto, M. (2013). Improved climate simulation by MIROC5: Mean states, variability, and climate sensitivity. *J. Climate*, *26*, 6312–6335. doi: 10.1175/2010JCLI3679.1

Wes McKinney. (2010). Data Structures for Statistical Computing in Python. In Stéfan van der Walt & Jarrod Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (p. 56 - 61). doi: 10.25080/Majora-92bf1922-00a

Wheeler, M. C., & Hendon, H. H. (2004). An All-Season Real-Time Multivariate MJO Index: Development of an Index for Monitoring and Prediction. *Monthly Weather Review*, *132*(8), 1917-1932. doi: 10.1175/1520-0493(2004)132⟨1917:AARMMI⟩2.0.CO;2

Wolter, K., & Timlin, M. S. (1993). Monitoring ENSO in COADS with a Seasonally Adjusted Principal Component Index. In *Proceedings of the 17th Climate Diagnostics Workshop* (Vol. 57, pp. 52–57).

Wolter, K., & Timlin, M. S. (1998). Measuring the strength of ENSO events: How does 1997/98 rank? *Weather*, *53*(9), 315-324. doi: 10.1002/j.1477-8696.1998.tb06408.x

Wolter, K., & Timlin, M. S. (2011). El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext). *International Journal of Climatology*, *31*(7), 1074-1087. doi: 10.1002/joc.2336

Wu, P. P.-Y., Julian Caley, M., Kendrick, G. A., McMahon, K., & Mengersen, K. (2018). Dynamic Bayesian network inferencing for non-homogeneous complex

systems. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *67*(2), 417-434. doi: 10.1111/rssc.12228

Yang, D., & Saenko, O. (2012). Ocean heat transport and its projected change in canesm2. *J. Climate*, *25*, 8148–8163. doi: 10.1175/JCLI-D-11-00715.1

Zhang, T., Hoell, A., Perlwitz, J., Eischeid, J., Murray, D., Hoerling, M., & Hamill, T. M. (2019). Towards Probabilistic Multivariate ENSO Monitoring. *Geophysical Research Letters*, *46*(17-18), 10532-10540. doi: 10.1029/2019GL083946

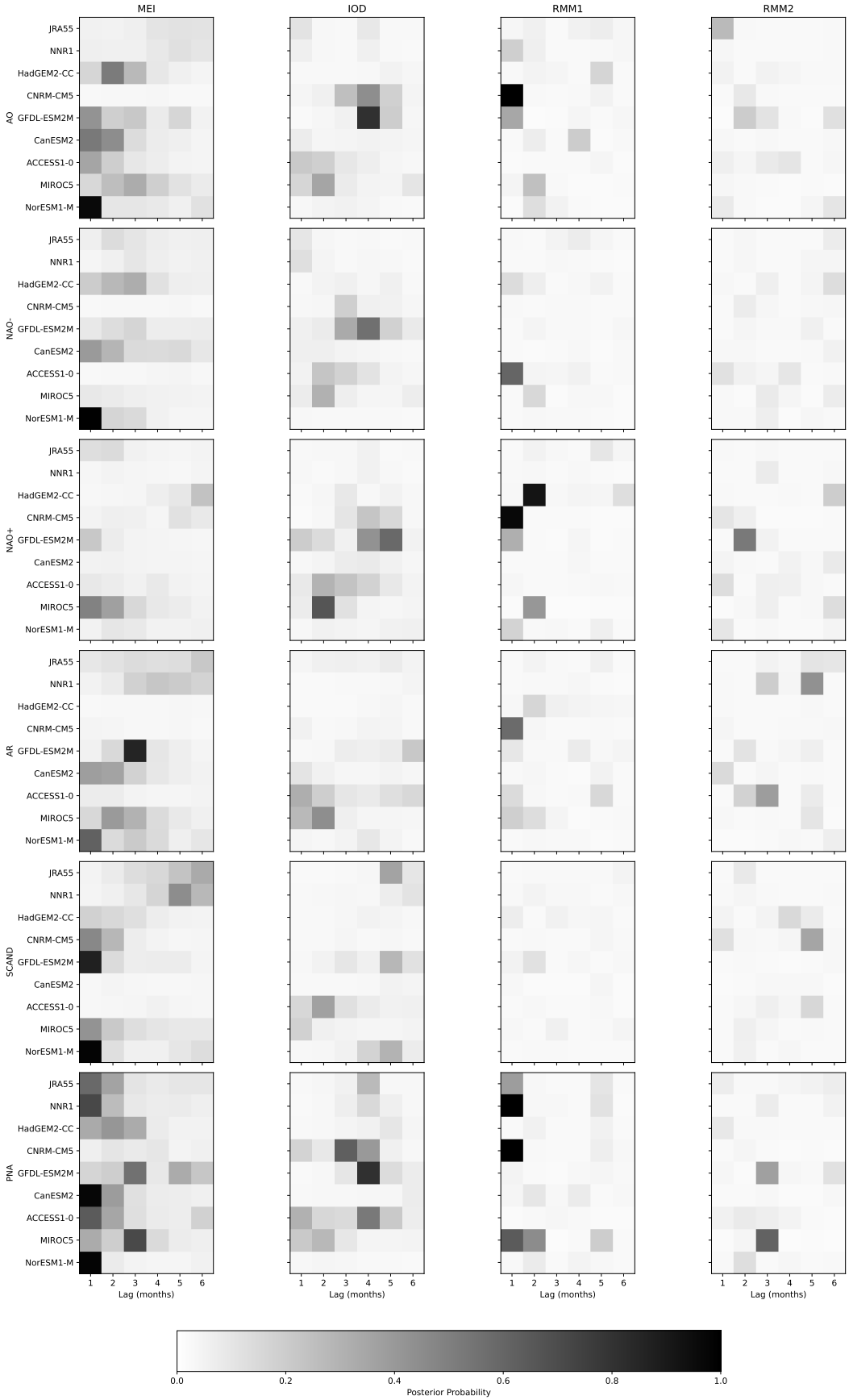

Figure 5. As for figure 4 but for the NH-tropical indices.

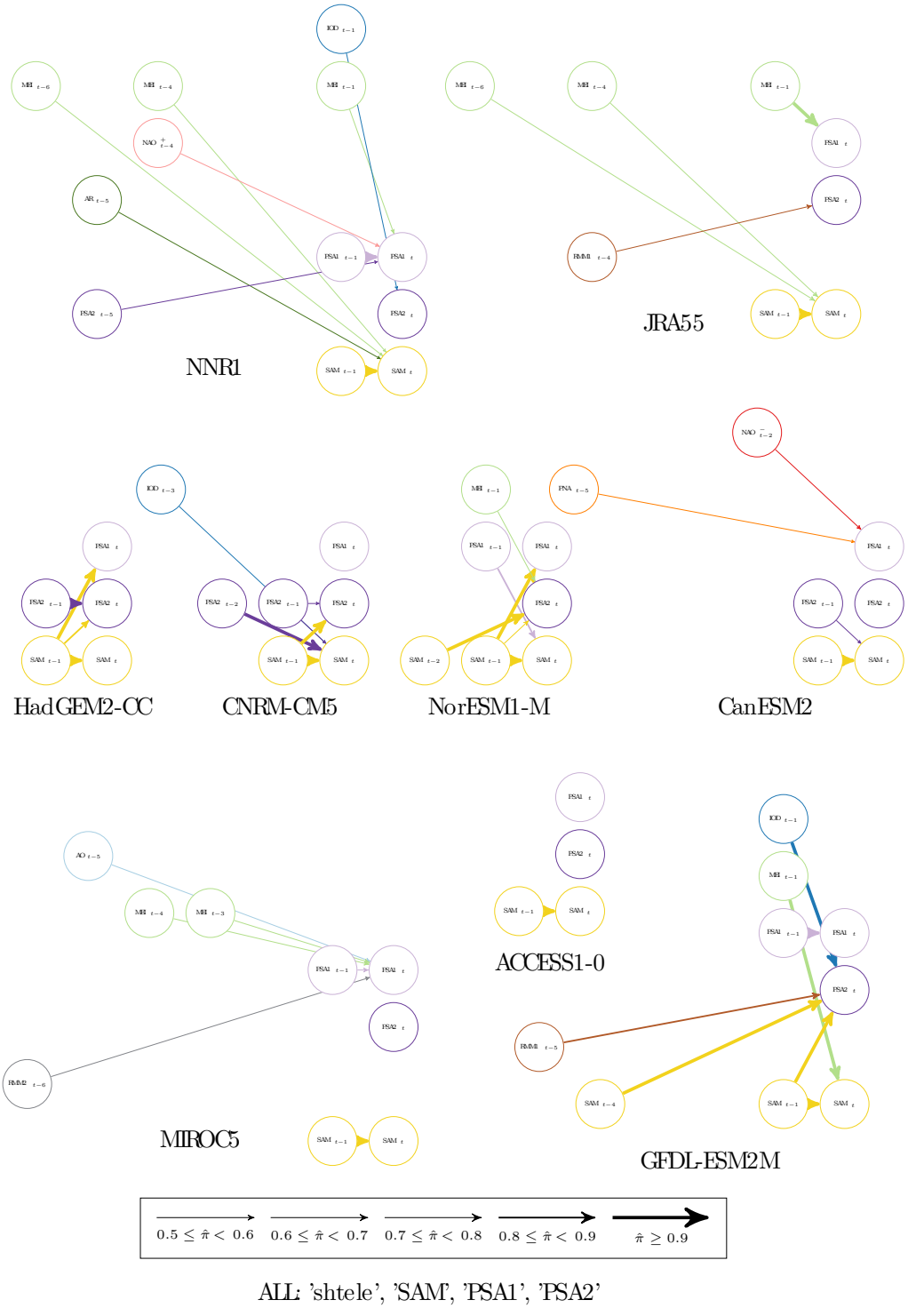

Figure 6. Edge posterior probabilties for the SH indices over all seasons (ALL: 'shtele') for the JRA-55 and NNR1 reanalyses and for the historical CMIP5 model simulations. Only edges with an estimated posterior weight greater than 0.5 are shown.

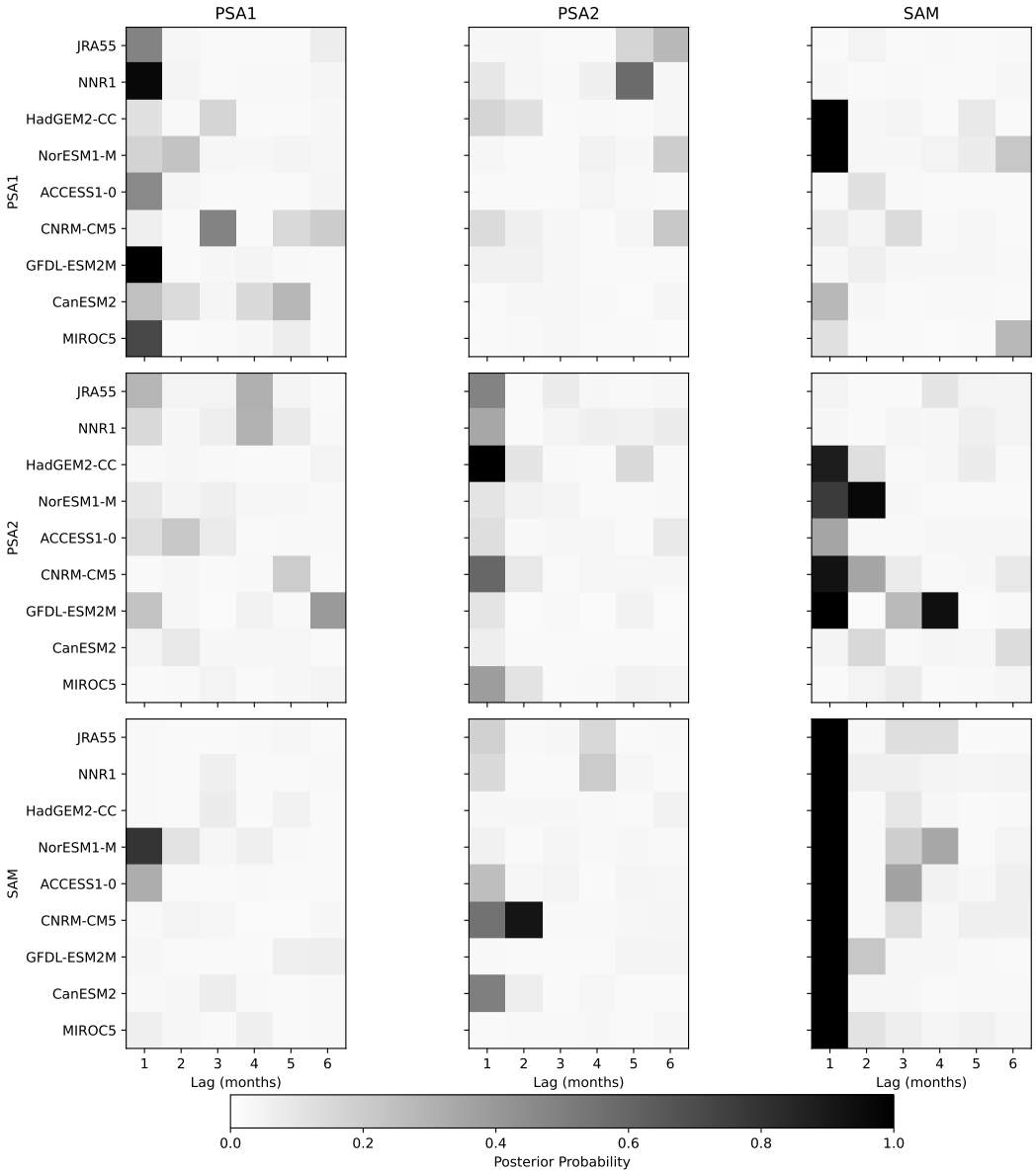

Figure 7. The heat map corresponding to the DAGs in figure 6.

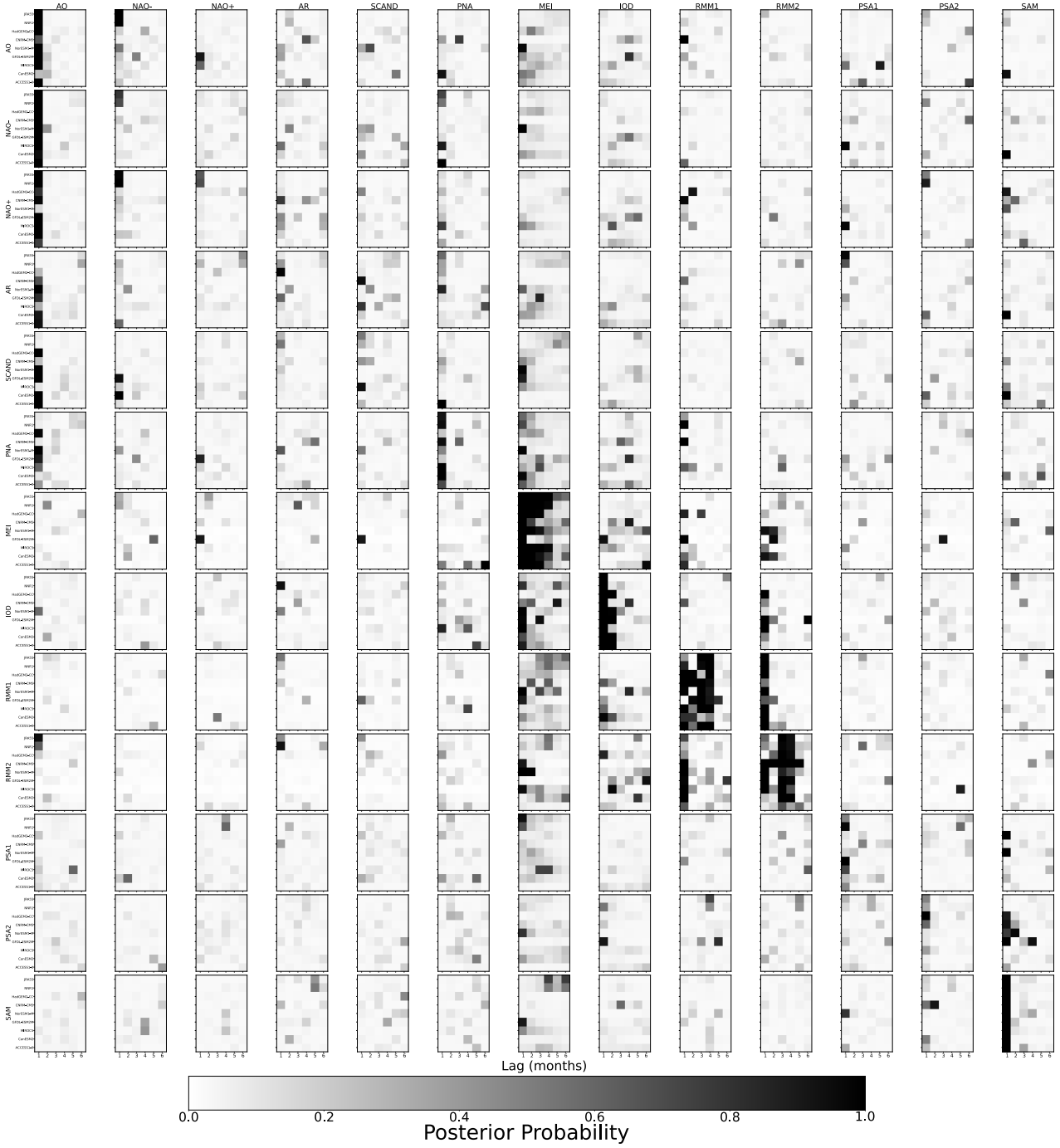

Figure 8. Heat map for the posterior weights i.e., MAP parent sets for monthly climate indices across the 7 CMIP models and 2 reanalyses calculated over all seasons. Here we show the estimated posterior probability $\hat{\pi}$ of the edge. The respective models shown in the rows are ordered based on their Wasserstein distance relative to the JRA-55 reanalysis.

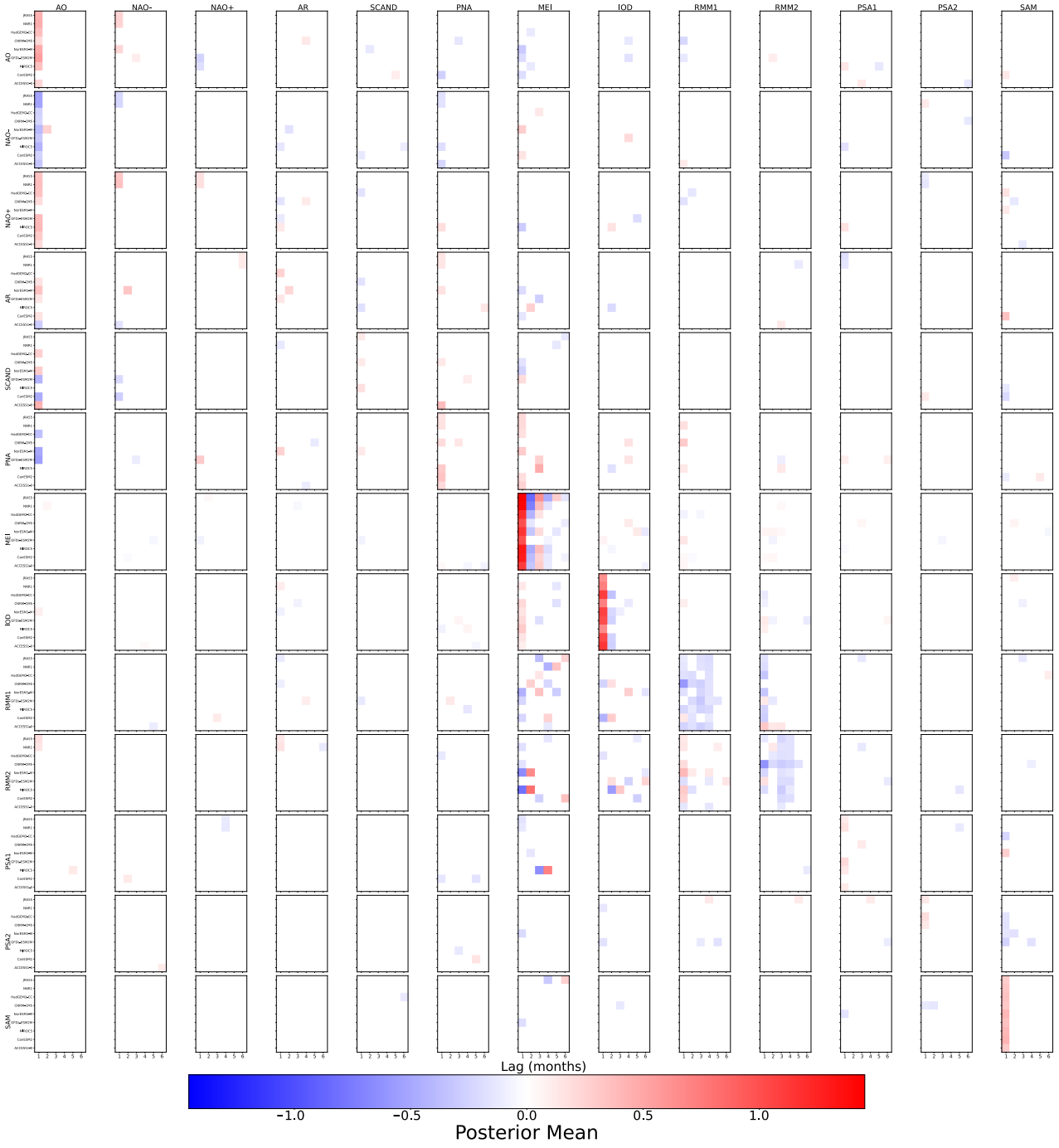

Figure 9. As for figure 8 but showing the posterior means i.e., mean parameter value $\hat{\beta}$ conditional on the MAP structure.

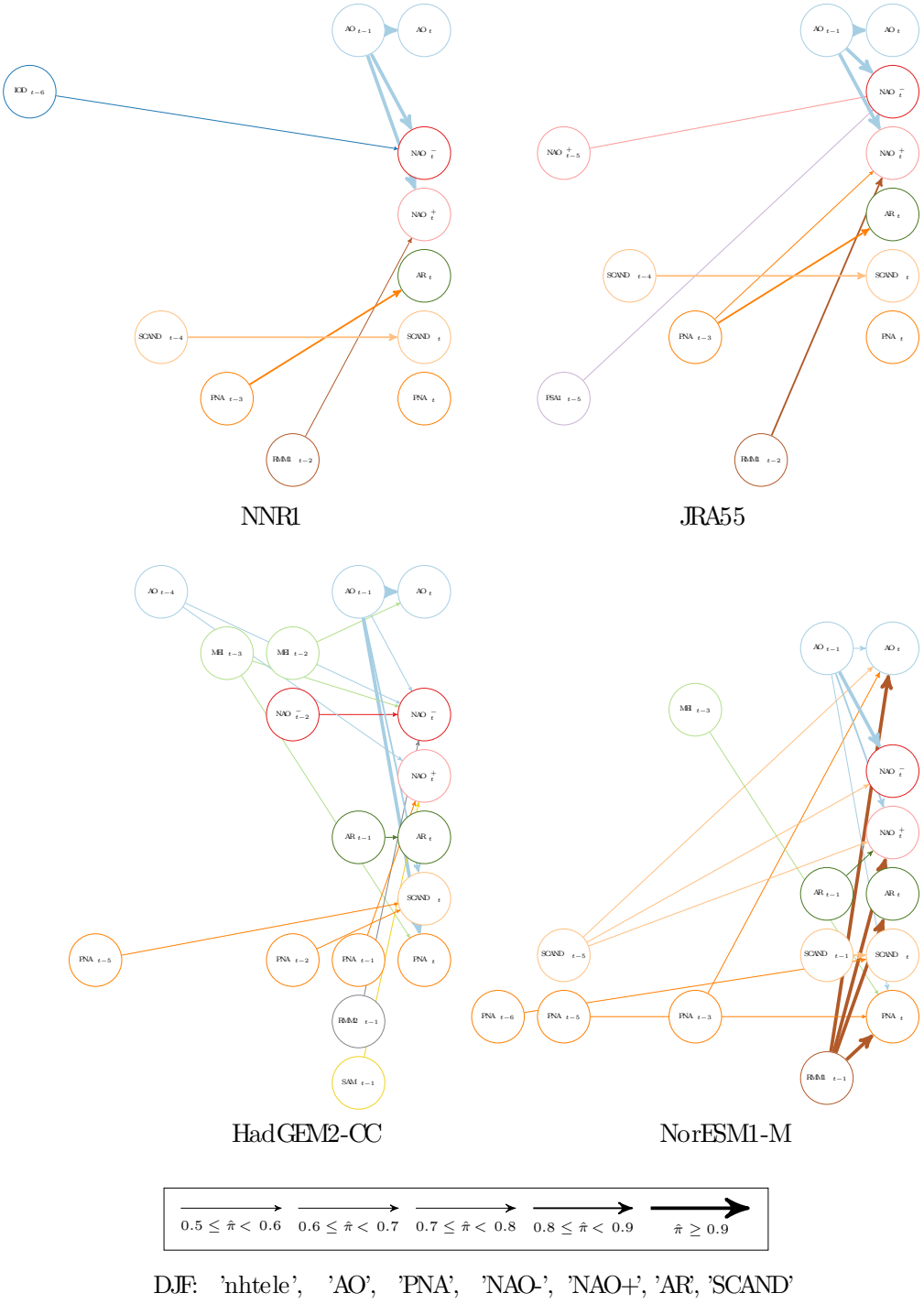

Figure 10. Edge posterior probabilties for the NH indices during the boreal winter (DJF: 'nhtele') for the JRA-55 and NNR1 reanalyses and for the HadGEM-CC, and NorESM1-M historical CMIP5 model simulations. Only edges with an estimated posterior weight greater than 0.5 are shown.

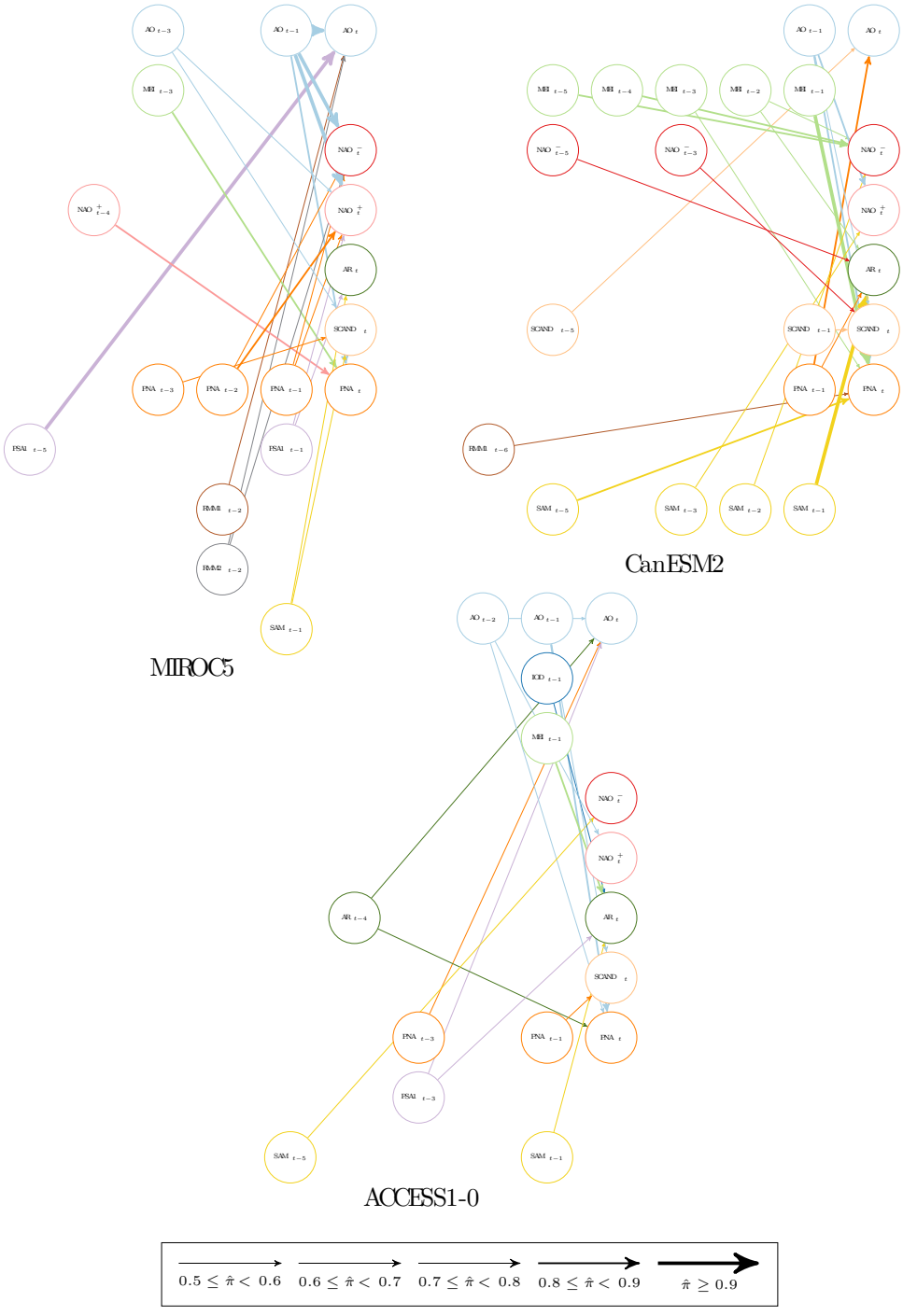

Figure 11. As for figure 10 but for MIROC5, CanESM2, and ACCESS1-0.

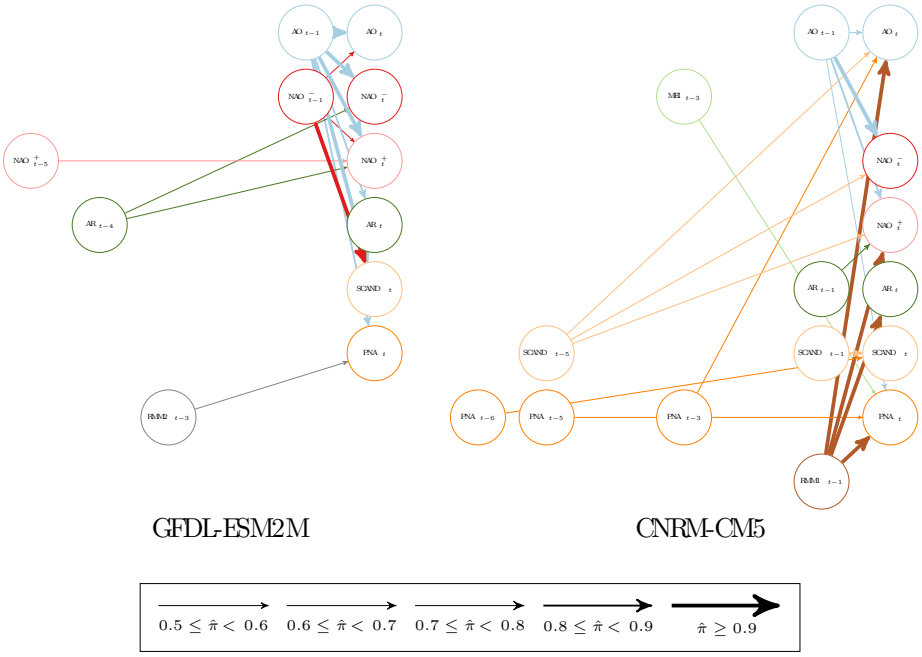

Figure 12. As for figure 10 but for GFDL-ESM2M, and CNRM-CM5.

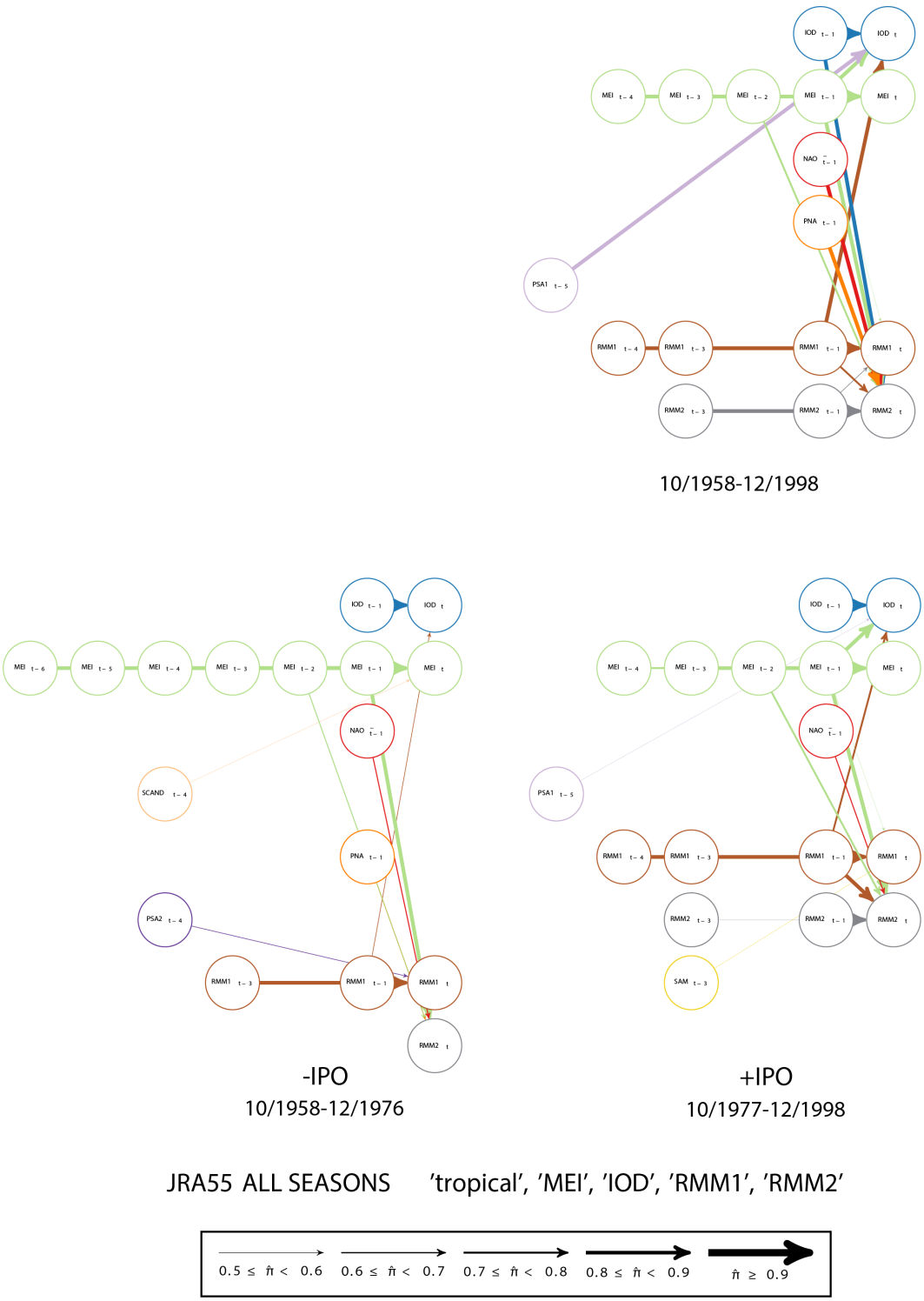

Figure 13. Tropical dependencies in the JRA-55 data over the period 1958-10-01 to 1998-12-31 encompassing two phases of the interdecadal Pacific oscillation (IPO) i.e., negative (-IPO: 1958-10-01 to 1976-12-31) and positive (+IPO: 1977-01-01 to 1998-12-31) IPO phases. Here all seasons are considered.