Posted on 23 Nov 2022 — CC-BY 4.0 — https://doi.org/10.1002/essoar.10505506.1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Dynamic Bayesian networks for evaluation of Granger causal relationships in climate reanalyses

Dylan Harries[1] and Terence John O'Kane[1]

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

November 23, 2022

Abstract

We apply a Bayesian structure learning approach to study interactions between global teleconnection modes, illustrating its use as a framework for developing process-based diagnostics with which to evaluate climate models. Homogeneous dynamic Bayesian network models are constructed for time series of empirical indices diagnosing the activity of major tropical, Northern and Southern Hemisphere modes in the NCEP/NCAR and JRA-55 reanalyses. The resulting probabilistic graphical models are comparable to Granger causal analyses that have recently been advocated. Reversible jump Markov Chain Monte Carlo is employed to provide a quantification of the uncertainty associated with the selection of a single network structure. In general, the models fitted from the NCEP/NCAR reanalysis and the JRA-55 reanalysis are found to exhibit broad agreement in terms of associations for which there is high posterior confidence. Differences between the two reanalyses are found that involve modes for which known biases are present or that may be attributed to seasonal effects, as well as for features that, while present in point estimates, have low overall posterior mass. We argue that the ability to incorporate such measures of confidence in structural features is a significant advantage provided by the Bayesian approach, as point estimates alone may understate the relevant uncertainties and yield less informative measures of differences between products when network-based approaches are used for model evaluation.

Dynamic Bayesian networks for evaluation of Granger causal relationships in climate reanalyses

Dylan Harries[1] and Terence J. O'Kane[1]

[1]CSIRO Oceans and Atmosphere, Hobart, Australia

Key Points:

- Bayesian structure learning provides a principled approach to quantifying uncertainty in estimated network structures for relationships between teleconnections
- Dynamic Bayesian networks estimated from NCEP/NCAR and JRA-55 reanalysis data show broad overall consistency
- Structural differences in high posterior credibility associations may be indicative of biases relevant for subsequent model evaluation

Corresponding author: Dylan Harries, `Dylan.Harries@csiro.au`

Abstract

We apply a Bayesian structure learning approach to study interactions between global teleconnection modes, illustrating its use as a framework for developing process-based diagnostics with which to evaluate climate models. Homogeneous dynamic Bayesian network models are constructed for time series of empirical indices diagnosing the activity of major tropical, Northern and Southern Hemisphere modes in the NCEP/NCAR and JRA-55 reanalyses. The resulting probabilistic graphical models are comparable to Granger causal analyses that have recently been advocated. Reversible jump Markov Chain Monte Carlo is employed to provide a quantification of the uncertainty associated with the selection of a single network structure. In general, the models fitted from the NCEP/NCAR reanalysis and the JRA-55 reanalysis are found to exhibit broad agreement in terms of associations for which there is high posterior confidence. Differences between the two reanalyses are found that involve modes for which known biases are present or that may be attributed to seasonal effects, as well as for features that, while present in point estimates, have low overall posterior mass. We argue that the ability to incorporate such measures of confidence in structural features is a significant advantage provided by the Bayesian approach, as point estimates alone may understate the relevant uncertainties and yield less informative measures of differences between products when network-based approaches are used for model evaluation.

Plain Language Summary

To produce reliable forecasts and projections, climate models should accurately reproduce the observed behavior of different processes that play a role in Earth's climate, including the relationships between them. Statistical methods can be used to describe these interactions in models and in observations, which can then be compared to evaluate how well a given model captures the observed relationships. However, networks obtained from estimates of the true historical state of the climate, known as reanalyses, will also be affected by the properties of the systems used to create these estimates, as well as random variability, and hence may have significant uncertainties. Using what are known as Bayesian statistical methods, we estimate the uncertainties associated with particular interactions in two widely used reanalyses. Interactions that are found to be very likely to be present in one reanalysis but not the other are suggested to be due to systematic differences in the two reanalysis systems and need to be kept in mind when these state

estimates are used to evaluate climate models. Therefore, it is important to account for the uncertainty associated with each relationship when analyzing state estimates and further employing them to evaluate climate models.

1 Introduction

The behavior of the Earth's climate system, from day-to-day changes in weather to longer-term variations in climate, arises as a result of the interactions of diverse processes within the coupled land-ocean-atmosphere-cryosphere system over a vast range of spatiotemporal scales (see, e.g., Ghil and Lucarini (2020) for a recent review). Developing an accurate understanding of the underlying processes, including their response to external forcing and the interactions between different components, is an important first step in the development of realistic numerical forecasting models. Inevitably, even with a good understanding of the main processes being simulated, any given model will still be limited in its ability to represent the climate system, e.g., due to deficiencies in the parameterization of unresolved processes. These limitations manifest as systematic biases in the output of the model compared to observations. By comparing the representation of a particular component in the model with its observed counterpart, shortcomings in the model implementation can be identified for improvement, at least subject to the severe limitation that the globally contiguous observational record of the climate system, and in particular the subsurface ocean, extends only over a few decades.

This process-oriented approach, in which attention is focused on a comparatively small set of physical processes, has been widely applied in the climate science community. On the one hand, the complexity of the full coupled climate system means that it is typically only feasible to focus on specific subsystems in any given analysis, or otherwise necessitates some form of dimension reduction. For instance, specific targeted process-oriented diagnostics (Maloney et al., 2019) permit the representation of these processes in models to be evaluated against observations in order to drive model development (Eyring et al., 2019). At the same time, the existence of distinct teleconnections, i.e., recurrent large-scale modes of variability, has long been recognized (Ångström, 1935), motivating simplified models of the climate in terms of a small number of interacting modes. For example, atmospheric teleconnection patterns such as the Arctic Oscillation (AO) (Thompson & Wallace, 1998), the North Atlantic Oscillation (NAO) (Walker, 1923; van Loon & Rogers, 1978), the Pacific North American (PNA) (Wallace & Gutzler, 1981; Horel & Wallace, 1981; Barnston & Livezey, 1987) and Pacific South American patterns (PSA) (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) constitute important sources of low-frequency variability (Hannachi et al., 2017) and are associated with wide-ranging impacts (see, e.g., Leathers et al., 1991; Thompson & Wallace, 1998; Mo & Paegle, 2001; Hurrell et al., 2003; Gillett et al., 2006). On interannual time-scales, modes that express in the tropical oceans such as the El Niño Southern Oscillation (ENSO) (Walker, 1924; Bjerknes, 1969) and the Indian Ocean Dipole (IOD) (Saji et al., 1999) emerge as dominant sources of variability, with important implications for global weather and climate (see, e.g., Schott et al., 2009; McPhaden et al., 2021). As these large-scale modes tend to be relatively persistent, understanding their evolution and dynamics may enable more skillful forecasting over longer time-scales (Goddard et al., 2001; A. G. Marshall et al., 2014; Hannachi et al., 2017). Thus, beyond simply being convenient for reducing the dimension of the problem in model evaluation studies, simplified models based on these physically observed modes provide an approach for better understanding the properties and interactions of key sources of climate variability.

Even when focusing on only a few individual processes, it is generally difficult to directly attribute model biases to problems in the representation of a single process (Eyring et al., 2019), or to understand the behavior of an individual teleconnection mode in isolation from other processes. To capture the intrinsically coupled nature of the system in simplified models, network based approaches have become increasingly popular (Tsonis & Roebber, 2004; Tsonis et al., 2006; Donges et al., 2009b; Steinhaeuser et al., 2011). In this framework, the climate system is represented in terms of a set of nodes, corresponding to appropriately defined subsystems or processes of interest, and edges describing the interactions between these nodes. The subsystems, for example, may be identified with one or more spatial gridpoints (Bello et al., 2015; Fountalis et al., 2018), or pre-defined modes characterized by empirical indices (Tsonis et al., 2007). In either case, the full climate system is then modeled in terms of a collection of individual, non-linear dynamical systems interacting with their neighbors in the constructed network. Such networks have variously been applied to study synchronization and climate shifts (Tsonis et al., 2007; G. Wang et al., 2009), to investigating the collective spatial structure of the statistical relationships between fields and changes over time (Tsonis & Swanson, 2008; Tsonis et al., 2008; Gozolchiani et al., 2008; Yamasaki et al., 2008; Donges et al., 2011;

Gozolchiani et al., 2011; Berezin et al., 2012; Guez et al., 2012; Steinhaeuser et al., 2012; Radebach et al., 2013; Y. Wang et al., 2013) and as tools for automatic mode identification (Bello et al., 2015) or dimension reduction (Fountalis et al., 2018; Falasca et al., 2019) via community detection methods.

When attempting to build a network representation of the climate, the structure of the network is generally not known beforehand and must be inferred by some means. A typical approach is to add or remove edges from the network on the basis of the level of statistical interdependence of pairs of nodes, quantified by, e.g., the correlation (Tsonis & Roebber, 2004) or mutual information (Donges et al., 2009a) between time series associated with each node. As these measures of association will almost always be non-zero in finite samples, some level of thresholding or pruning must also be applied in order to exclude edges corresponding to weak or spurious associations. In the simplest case, the result is an undirected network; that is, the presence of an edge between two nodes indicates some level of mutual association, but does not provide information on any possible directionality in the relationship.

Networks constructed in this way have proven to be very useful but do have some important limitations. In particular, while correlation graphs allow for comparisons between modeled and observed associations (Falasca et al., 2019), often it is of interest whether a particular set of variables is causally related to another set. If there is a causal mechanism, we may also wish to quantify the magnitude of the effects of those causal factors; commonly, climate networks based on the above approaches do not provide direct access to measures of effect. In practice, to answer these sorts of questions it is necessary to imbue the networks with additional structure. This can be naturally achieved by identifying the original graph with an underlying statistical model, that is, by working in the context of a (probabilistic) graphical model (Koller & Friedman, 2009). In a graphical model, the graph encodes, in a well-defined way, the set of qualitative independence relationships between the random variables, corresponding to nodes, in the model (Jordan, 2004). Given a particular functional form for the interactions between variables in terms of the joint probability density function (PDF), quantitative questions can also be formulated and addressed using standard algorithms (Pearl, 1982, 1988; Dechter, 1999; Koller & Friedman, 2009).

Graphical models may utilize undirected or directed edges, or even a mixture of both, corresponding to different ways of constructing the joint PDF of the model. Models based on directed acyclic graphs (DAGs), in which all edges also have an associated direction and the graph does not contain any directed closed loops, provide an intuitive representation of complex systems in terms of conditional independence relationships between quantities (Pearl, 1988; Spiegelhalter et al., 1993; Jordan, 2004). Compared to undirected networks, directed graphs have the advantage that edges in the graph can often be interpreted causally. As a result, they provide a powerful tool for studying causal relationships (Pearl, 1995), and form a useful basis for other forms of causal inference (Greenland & Brumback, 2002). Bayesian, or belief, networks (BNs), as such models are usually called, have therefore received increasing attention from the climate science community (Ebert-Uphoff & Deng, 2012a), having variously been used for forecasting and risk assessment based on expert systems (e.g., Abramson et al., 1996; Catenacci & Giupponi, 2009; Peter et al., 2009; Catenacci & Giupponi, 2013; Leonard et al., 2014; Boneh et al., 2015), for learning independence relationships and possible causal interactions in observations (Ebert-Uphoff & Deng, 2012b; Runge et al., 2014; Runge, 2015; Runge et al., 2015; Kretschmer et al., 2016, 2017; Horenko et al., 2017; Li et al., 2018; Runge, 2018a, 2018b; Runge, Nowack, et al., 2019; Runge, Bathiany, et al., 2019; Samarasinghe et al., 2019, 2020; Saggioro et al., 2020; Pfleiderer et al., 2020; Di Capua et al., 2020) and models (Deng & Ebert-Uphoff, 2014; Ebert-Uphoff & Deng, 2017), and, most recently, for model evaluation (Vázquez-Patiño et al., 2020; Nowack et al., 2020).

One approach for using BNs as tools for model evaluation is to learn the network structure in a model and in observations and assess the agreement between the two. Learning the structure of climate networks, in the absence of expert knowledge, has primarily been achieved by utilizing constraint-based algorithms (Spirtes & Glymour, 1991; Colombo & Maathuis, 2014) in which the set of edges is determined starting from a series of conditional independence tests (Ebert-Uphoff & Deng, 2012a; Runge, Bathiany, et al., 2019). Constraint-based methods can flexibly incorporate linear or non-linear conditional independence tests (Hlinka et al., 2013) together with predefined constraints, allowing for non-linear dependence structures to be estimated from data. However, the inclusion of edges on the basis of an (initially arbitrary) significance level together with multiple testing adjustments makes assessing the level of confidence in the inferred networks difficult.

Usually, sensitivity analyses are necessary to get some handle on the robustness of the resulting networks.

Ideally though, to apply BNs as tools for model evaluation we would not only identify differences between modeled and observed networks, but also assess whether any mismatch is likely due to a model bias or simply sampling variability. In other words, an additional level of uncertainty quantification is required. While a great deal of insight can be derived from the structure of the estimated networks, quantifying both the sign and magnitude of the interaction between nodes has generally been achieved by performing a second stage of fitting a parametric model, conditional on the inferred structure. This, again, leads to complications in determining the uncertainties in the estimated interaction strengths, which (depending on the approach used) may be insufficiently conservative (Draper, 1995). For the purposes of model evaluation, it is usually of interest whether a model captures both the existence of a link and with the correct strength. Unreliable estimates for the uncertainty in fitted interaction strengths may result in an inability to determine if an interaction is present but differs significantly in the model compared to observations. On top of this, the additional complexity involved in implementing such two-stage fitting procedures appears to have discouraged their use (McGraw & Barnes, 2018) compared to simpler model-based analyses framed in terms of Granger causality (Granger, 1969).

In this article, we investigate possible solutions to the above limitations through the use of Bayesian methods (Uusitalo, 2007). In the Bayesian framework for learning the structure and effect measures (Spiegelhalter & Lauritzen, 1990; Dawid & Lauritzen, 1993; Madigan et al., 1995), it is natural to use the posterior probability of a given network and its associated parameters as a score to measure model fitness (Buntine, 1991; Cooper & Herskovits, 1992; Geiger & Heckerman, 1994; Heckerman et al., 1995). Existing knowledge and constraints may be incorporated through the use of suitable prior distributions, although in practice this must be balanced against computational feasibility. Bootstrap (Friedman et al., 1999) or sampling-based (Madigan & Raftery, 1994; Madigan et al., 1995; Godsill, 2001) methods provide some measure of the uncertainties in model selection, as well as allowing the predictions of multiple models to be combined via model averaging (Madigan & Raftery, 1994). Consequently, the Bayesian approach provides a principled quantification of the uncertainties associated with estimation of the network structure and parameters. Models typically used for testing for Granger causal relationships, that is, linear autoregressive models, can be straightforwardly expressed in graphical terms (Arnold et al., 2007; Lèbre, 2009), and, at least under a choice of conjugate priors, analyzed efficiently using closed-form expressions for the desired posterior densities (Geiger & Heckerman, 1994). In this sense, the model specification is largely familiar, reducing the barriers to use in climate applications. While there is additional complexity associated with the inference scheme, we believe that this is justified by the need for some level of uncertainty quantification.

To illustrate the utility of Bayesian methods for structure learning, we consider the application of fitting BNs to a set of teleconnection indices derived from two reanalysis datasets. Our purpose in doing so is two-fold. Firstly, as the score-based approach has not been widely used in analyses of this type, we wish to investigate the suitability of the method for networks of the size encountered in realistic data. Our second, and more important, aim is to perform a comparison of different reanalysis products on the basis of the fitted networks. In addition to allowing for possible biases and differences in the reanalyses at the level of interactions between modes to be studied, differences in the networks from different reanalyses give some additional indication of uncertainties in the observed networks that model runs are evaluated against. This, in turn, must be taken into account when deciding what level of disagreement between models and observations can be taken to indicate clear model biases.

The remainder of this paper is structured as follows. In the next section we provide a brief review of Bayesian network models and the inference methods used to fit such models. In Section 3 we describe the datasets and diagnostics that we study, together with our choice of prior distributions. In Section 4 we present the results of fitting the network models. Finally, we summarize our findings and discuss their implications for follow-up comparisons in Section 5.

2 Dynamic Bayesian networks

2.1 Structure learning

As noted above, a graphical model is simply a statistical model that has associated with it a graph encoding the relationships between the variables in the model. Each random variable in the model is represented by a node in the graph, with the allowed conditional dependence relationships between variables indicated by edges between nodes.

Under suitable assumptions (see, e.g., Koller and Friedman (2009) for a review), translating between graphs and joint PDFs that constitute a model can be achieved using a prescribed set of rules.

For BNs and other graphical models based on DAGs, the graph structure implies a factorization of the joint PDF into conditional density functions. Given a set of random variables $Y^1, \ldots, Y^n$ and a DAG G, we denote by $\mathrm{pa}_G(Y^i)$ the set of nodes in G that have a directed edge connecting to Y^i,

$$\mathrm{pa}_G(Y^i) = \{Y^j | G \text{ contains an edge from } Y^j \text{ to } Y^i\}. \tag{1}$$

The graph G then provides a representation of the joint PDF $P(Y^1, \ldots, Y^n)$ if the PDF admits a factorization of the form

$$P(Y^1, \ldots, Y^n) = \prod_{i=1}^{n} P(Y^i | \mathrm{pa}_G(Y^i), \theta_i), \tag{2}$$

where θ_i denotes any parameters required to characterize the conditional density. For example, when all of the variables in $\{Y^i\} \cup \mathrm{pa}_G(Y^i)$ are discrete, the conditional density $P(Y^i | \mathrm{pa}_G(Y^i), \theta_i)$ is the conditional probability table (CPT) summarizing the probability of observing each level of Y^i for each combination of values of the parents $\mathrm{pa}_G(Y^i)$; the θ_i are simply the values of each probability in the table. For continuous variables, the θ_i are any parameters required to fully specify the corresponding continuous PDF.

Generally in geophysical applications, the random variables of interest exhibit nontrivial spatial and temporal correlations. In our case, these variables are a collection of (continuous) teleconnection indices, some of which (e.g., ENSO) show substantial autocorrelation. Feedback loops or temporal dependence of this form cannot be represented in a BN with a single node for each index Y^i, due to the requirement for the graph G to be acyclic (Uusitalo, 2007). To handle this, the set of nodes is expanded to consist of the values of the random variables Y^i at the current time t, Y^i_t, as well as the values of the variables $Y^i_{t-\tau}$ 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 $\tau_{\max}$. Temporal dependencies are described by edges between the nodes corresponding to, say, Y^i_t and $Y^i_{t-\tau}$, and a full time series of observations is described by a graph at each point in time relating the values of the variables at that time point to those in the previous time-slices. Similar graphical models for multivariate time series have also been introduced as time series graphs (Eichler, 2012). The resulting model is referred to as a dynamic Bayesian network (DBN).

In the simplest case, the structure of the graph and the associated parameters remains the same across all time-slices, so that the full time series is modeled by repeating the graph at each time t. The corresponding DBN is said to be (time-)homogeneous (K. P. Murphy & Russell, 2002). The assumption of homogeneity is often violated in geophysical applications, however. Interactions between spatially separated processes, for example, may be seasonally dependent, while on longer time-scales the possibility of climate regime changes, e.g., in association with tipping points (Lenton et al., 2008), in response to anthropogenic forcing has recently become a key concern. Non-homogeneous DBNs, in which either the graph structure, parameters or both simultaneously, are allowed to change over time, admit the possibility of modeling features such as secular trends and regime changes (Wu et al., 2018), at the cost of a significant increase in complexity in terms of model specification and inference. Here we focus on the simpler case of homogeneous models for the purposes of investigating the usefulness of Bayesian methods for assessing model uncertainty; the more complicated case of non-homogeneous models will be described in a separate study.

Fitting a homogeneous DBN to an observed time series $D = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$, 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, requires learning (in general) both the structure of the graph G and the values of the corresponding parameters. Since

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

where θ denotes the collection of all parameters for the conditional PDFs, the learning process can be conveniently divided into two steps. In the first, structure learning stage, the structure of the graph G is sought, independent of specific values of the parameters. Structure learning methods for BNs can be roughly categorized as constraint-based or score-based. The former set of methods attempt to reconstruct the graph structure on the basis of conditional independence tests and available prior knowledge and constraints, using, e.g., the PC-algorithm (Spirtes & Glymour, 1991) and its extensions (e.g., Colombo & Maathuis, 2014; Runge, 2018a; Runge, Nowack, et al., 2019). As discussed in Section 1, most recent examples in climate science have made use of constraint-based algorithms to learn an initial structure, followed in some cases by a separate parameter learning step. In contrast, in score-based approaches the graph G is estimated based on maximizing a suitable score function (Cooper & Herskovits, 1992; Geiger & Heckerman, 1994; Heckerman et al., 1995), such as the marginal likelihood $P(D|G)$ or an information criterion.

However, when there may be significant uncertainty associated with the selection of a single model, rather than finding a single optimal model, it may be preferable to attempt to account for this model uncertainty by sampling from the full posterior distribution of possible graphs $P(G|D)$ (Madigan et al., 1995). Estimates for derived quantities of interest Δ may then be computed by averaging over the posterior distribution (Madigan & Raftery, 1994; Draper, 1995),

$$\Pr(\Delta|D) = \sum_{G \in \mathcal{G}} \Pr(\Delta|G, D) P(G|D) \approx \frac{1}{S} \sum_{s=1}^{S} \Pr(\Delta|G^{(s)}, D), \tag{4}$$

where $\mathcal{G}$ is the space of allowed structures and $\{G^{(s)}\}_{s=1}^{S}$ is a sample of size S from the posterior distribution $P(G|D)$. In particular, structural uncertainties may be quantified by taking Δ to be an indicator function for the presence of a given edge, with Eq. (4) quantifying the posterior probability of the existence of that edge, given the chosen model class and observed data.

Directly sampling the posterior $P(G|D)$ can be achieved in the case that the marginal likelihood

$$P(D|G) = \int d\theta P(D|G, \theta) P(\theta|G) \tag{5}$$

can be evaluated, where $P(\theta|G)$ denotes a set of priors for the full set of node PDF 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,

$$P(D|G, \theta) = \prod_{t=1}^{T} \prod_{i=1}^{n} P(Y_t^i | \mathrm{pa}_G(Y_t^i), \theta_i); \tag{6}$$

note that the inner multiplication follows from the assumption that the joint PDF can be factored according to G, and the outer multiplication from the assumption of homogeneity. For simplicity, we work with the conditional likelihood assuming a sufficiently large set of pre-sample values are available to condition on. We restrict our attention to structures in which the parent set of a variable Y_t^i is not allowed to contain variables at the same time t, that is, we do not allow contemporaneous dependencies among variables. Excluding models with instantaneous links ensures that the structures we allow naturally satisfy structural modularity, such that the parent set of a variable Y_t^i may be chosen independently of the parent set for any other variable (Friedman & Koller, 2003). However, it should be kept in mind that doing so prevents any interactions that occur on time-scales that are shorter than the data sampling frequency from being directly handled, as these would manifest as instantaneous dependencies or feedback loops at $\tau =$

0. Assuming that the priors $P(\theta|G)$ satisfy the properties of parameter independence (Heckerman et al., 1995),

$$P(\theta|G) = \prod_{i=1}^{n} P(\theta_i|G), \tag{7}$$

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 $P(\theta_i|G) = P(\theta_i|G')$), the marginal likelihood may be written as the product of local marginal likelihoods $\Psi_i(D, G)$ (Grzegorczyk & Husmeier, 2011):

$$P(D|G) = \prod_{i=1}^{n} \int d\theta_i \prod_{t=1}^{T} P(Y_t^i|\mathrm{pa}_G(Y_t^i), \theta_i) P(\theta_i|G) \equiv \prod_{i=1}^{n} \Psi_i(D; G). \tag{8}$$

For structurally modular priors of the form

$$P(G) = \prod_{i=1}^{n} P(\mathrm{pa}_G(Y_t^i) \tag{9}$$

the posterior over graphs also factorizes,

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} = \frac{1}{P(D)} \prod_{i=1}^{n} \Psi_i(D; G) P(\mathrm{pa}_G(Y_t^i)), \tag{10}$$

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

2.2 Choice of conditional densities

The conditional densities $P(Y^i|\mathrm{pa}_G(Y^i), \theta_i)$ can, in principle, be chosen to model arbitrary relationships between the random variables in the graph, consistent with the independence assumptions embodied by the graph structure. In practice, this prevents marginalizing out the graph parameters (i.e., evaluating $\Psi_i(D; G)$ analytically) to sample from the marginal posterior distribution $P(G|D)$ directly. It is then necessary to construct a Markov Chain Monte Carlo (MCMC) sampler that samples from the joint posterior $P(\theta, G|D)$ using, e.g., reversible jump MCMC (RJMCMC) (Green, 1995) or related methods (Carlin & Chib, 1995; Godsill, 2001). A summary of samplers that we use is given in Appendix A.

In special cases, the necessary integrals can be evaluated in closed form, allowing for the posterior $P(G|D)$ to be efficiently sampled after marginalizing out the conditional PDF parameters. For continuous data, a widely used example for which this is possible is the linear Gaussian regression model (Punskaya et al., 2002; Lèbre et al., 2010). In this model, which can be regarded as a specialization of the BGe model for continuous data (Geiger & Heckerman, 1994), each Y_t^i is assumed to be conditionally Gaussian distributed,

$$Y_t^i | \mathrm{pa}_G(Y_t^i), \tau_i^2 \sim N(\mu_t^i, \tau_i^{-2}),$$
$$\mu_t^i = \beta_0^i + \sum_{j=1}^{p_i} \beta_{(k_j,\tau_j)}^i Y_{t-\tau_j}^{k_j} \tag{11}$$

with mean $\mu_t^i = E[Y_t^i | \mathrm{pa}_G(Y_t^i)]$ given by a linear function of the parent variables $\mathrm{pa}_G(Y_t^i) = \{Y_{t-\tau_j}^{k_j} | j = 1, \dots, p_i\}$. The local marginal likelihoods $\Psi_i(D; G)$ and posterior distributions for the parameters of a given graph can be analytically evaluated provided that conjugate normal-gamma priors are assumed for the conditional precision τ_i^2 and coefficients $\beta_{(k_j,\tau_j)}^i$,

$$\tau_i^2 \sim \mathrm{Gamma}(a_\tau, b_\tau),$$
$$\beta_{(k_j,\tau_j)}^i | \tau_i^2, \mathrm{pa}_G(Y_t^i) \sim N\left(0, \frac{\nu_i^2}{\tau_i^2}\right), \quad j = 1, \dots, p_i, \tag{12}$$

where a_τ, b_τ, and ν_i^2 are prior hyperparameters. Similar linear models have previously been applied (Kretschmer et al., 2016, 2017; Saggioro et al., 2020; Di Capua et al., 2020) to perform estimation of the interaction strengths after an initial stage of constraint-based structure learning. Using the underlying generative model, Eq. (11), posterior estimates for both the structural features and model parameters can be obtained within a single sampling scheme. Appropriate choices for the hyperparameters a_τ, b_τ, and ν^2 allow varying levels of regularization to be imposed so as to yield more reliable, if more conservative, estimates given relatively short and noisy time series. Alternatively, they may be allowed to vary and another level of priors specified for the unknown hyperparameters.

In the presence of significant non-linearity, linear models of this form may no longer be appropriate. In this case, one strategy to capturing the underlying non-linear relationship is to first discretize the original data and employ models for discrete data, e.g., the analytically tractable BDe model (Buntine, 1991; Cooper & Herskovits, 1992; Heckerman et al., 1995). However, this comes at the cost of necessarily losing some information, generally yielding only a coarse approximation of the original continuous distribution (Friedman & Goldszmidt, 1996). Additionally, choosing an appropriate discretization scheme is in general difficult, as there is an inevitable trade-off between having sufficient resolution to describe the data versus the exponential growth in the number of parameters that are required to specify the CPT of each variable. Hence, in the following we focus on the case of continuous data.

For a given choice of model, after performing any initial pre-processing, fitting the homogeneous DBN model to the observed indices data then consists of sampling from the posterior distribution $P(G|D)$ using the known marginal likelihood $P(D|G)$ in order to derive a set of candidate networks. For a given graph drawn from $P(G|D)$, the posterior distribution for the parameters, and hence summary statistics and credible intervals, can be computed analytically or obtained via standard within-model MCMC methods.

2.3 Toy model example

To make the above discussion more concrete, it is helpful to consider a simple toy example. A standard problem in climate studies is to determine the direction of the relationships, if any, between some set of variables. For example, during the positive (El Niño) phase of ENSO, anomalously warm Pacific sea surface temperatures (SSTs) drive elevated mean surface temperatures over North and South America (McGraw & Barnes, 2018). As a simplified example of this sort of driver-response relationship, McGraw and Barnes (2018) considered a two-dimensional system consisting of two observables D_t and R_t that evolve according to

$$\begin{aligned} D_t &= \alpha D_{t-1} + \sqrt{1-\alpha^2}\epsilon_t^D, \\ R_t &= D_{t-\tau} + \gamma\epsilon_t^R. \end{aligned} \tag{13}$$

The innovations ϵ_t^D and ϵ_t^R are taken to be independent Gaussian noise drawn from a standard normal distribution. Typically in climate analyses, the relationships between the system variables would be studied by regressing the postulated response (e.g., surface temperature anomalies or R_t) on lagged values of the driver (SST anomalies or D_t),

$$R_t = c_0 + \sum_{j=1}^{k} c_j D_{t-j}, \tag{14}$$

and vice versa to test for the possibility of the reversed relationship. However, this approach is susceptible to detecting spurious relationships when one or both processes exhibit substantial autocorrelation. As noted above, this often occurs in climate applications, where, e.g., the driver may correspond to relatively slowly varying boundary conditions such as SST driving an atmospheric response, as in the ENSO-surface temperature example. To account for this, autoregressive models of the form

$$R_t = a_0 + \sum_{j=1}^{k} a_j R_{t-j} + \sum_{j=1}^{k} b_j D_{t-j}, \tag{15}$$

(and similarly for the dependence of D_t on R_t), may be used instead to test for the presence of Granger causal links between processes.

Depending on the system, however, there can be considerable uncertainty associated with selecting one of these models over the other. We can apply the sampling based approach described in the previous sections to better quantify this. The two models, Eq. (14) and Eq. (15), correspond to particular choices of parent set under the linear Gaussian regression model. Thus, to illustrate the method we consider the results of learning the structure of a DBN describing the system Eq. (13) under a linear Gaussian model. For given values of the system parameters, we generate a random realization of the system and fit a DBN after standardizing the input data to zero mean and unit variance. For the prior hyperparameters, we take, for example, $a_\tau = 1.5$, $b_\tau = 10$, and $\nu^2 \approx 43.3$, yielding a weakly informative t prior with 3 degrees of freedom for each coefficient, with 90% prior credible intervals of $-4 \leq \beta \leq 4$. The prior 1% and 99% percentiles for the conditional precision (variance) are 0.57 (0.02) and 56.7 (1.74), respectively. Alternative choices with a heavier tailed distribution for the coefficients and much broader priors for the conditional precision, e.g., $a_\tau = 0.5$, $b_\tau = 10$, and $\nu^2 \approx 2$, tend to yield similar parameter estimates but less sparse models. The set of models considered includes all lags of up to 6 time steps and at most 4 parent nodes for each variable. Posterior samples are obtained using the MC^3 algorithm described in Appendix A, with the choices for the structure priors and proposal densities as described in Section 3.3.

Graphical summaries of the results of fitting this Gaussian DBN to realizations of the system Eq. (13) for $(\alpha, \gamma, \tau) = (0.2, 1, 2)$, $(0.5, 10, 2)$, and $(0.9, 4, 2)$ are shown in Figure 1. The true system, Eq. (13), corresponds to the graph given in Figure 1(a), with the edges labeled by the values of the standardized regression coefficients. The remaining panels show the estimated posterior probabilities $\hat{\pi}$ for each edge, computed as in Eq. (4), for each realization of the system; for clarity, only edges for which $\hat{\pi} > 0.5$ are shown. Where an edge also appears in the maximum a posteriori (MAP) estimate for the structure, the posterior 95% highest density interval (HDI) is also shown for the corresponding coefficient.

For low levels of autocorrelation and noise, the model recovers the correct edges with virtual certainty; in this case, the MAP structure consists of the (true) parent sets $\text{pa}(D_t) = \{D_{t-1}\}$ and $\text{pa}(R_t) = \{D_{t-2}\}$. Recovery of the true dependence structure is more difficult in the presence of large amounts of noise (e.g., $\gamma = 10$) or high autocorrelation (e.g., $\alpha = 0.9$). McGraw and Barnes (2018) note that, in the latter case, a bivariate Granger causality analysis exhibits reduced power for detecting the $D_{t-\tau} \rightarrow$

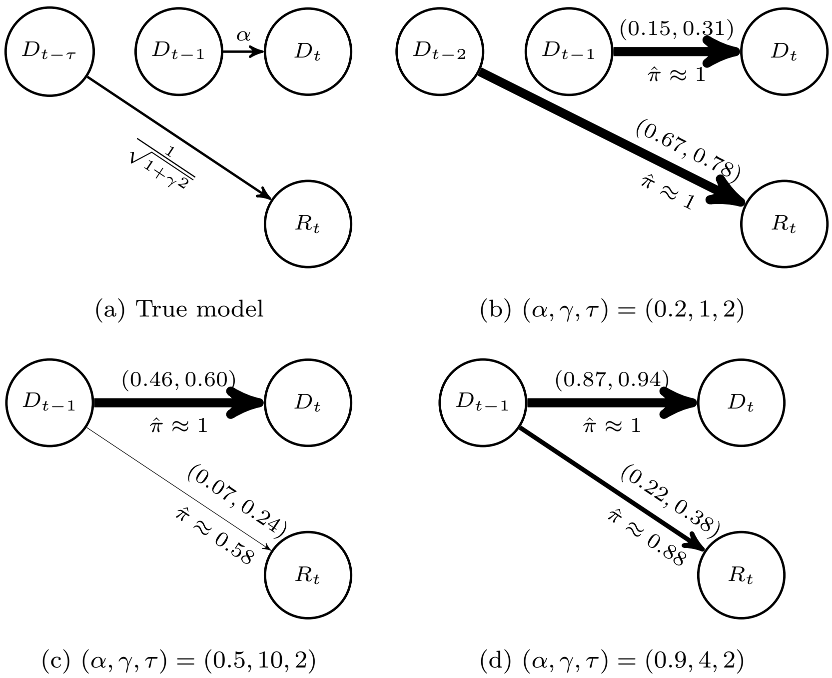


Figure 1. DBNs corresponding to (a) the true model given by Eq. (13), and network structures inferred using a linear Gaussian model for a single realization of the process with (b) $\alpha = 0.2$, $\gamma = 1$, $\tau = 2$; (c) $\alpha = 0.5$, $\gamma = 10$, $\tau = 2$; and (d) $\alpha = 0.9$, $\gamma = 4$, $\tau = 2$. For each of the fitted networks, all edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown, with thickness indicating the value of $\hat{\pi}$. Where an edge also occurs in the MAP model, the approximate posterior 95% HDI for the corresponding coefficient is also shown.

R_t edge while lagged regressions tend to too frequently identify spurious wrong-way relationships in which R drives D. For high levels of noise, both methods show reduced power for detecting the driver-response relationship, although accounting for autocorrelation successfully controls the false-positive rate for the reversed relationship. This difficulty in identifying the true underlying model is clearly evident in the lower panels of Figure 1 from the reduced estimated posterior probability of a dependence of R_t on D_t. In both cases, although the MAP structure suggests that R_t depends on lagged values of D_t, the lag at which this occurs is misidentified, with $\mathrm{pa}(R_t) = \{D_{t-1}\}$. Across all sampled models, the estimated posterior probability for lagged values of D_t to R_t reflects greater uncertainty in the model structure, and hence in the selection of a single optimal model. For instance, while the MAP parent set for R_t when $\alpha = 0.5$, $\gamma = 10$ is found to contain only D_{t-1}, the next most likely parent sets (for this particular sample, $\mathrm{pa}(R_t) = \emptyset$ and $\mathrm{pa}(R_t) = \{D_{t-2}\}$) also have non-negligible posterior probabilities, so that in practice it may be important to take this model uncertainty into account. A similar uncertainty measure is considerably more difficult to construct and interpret in the testing based approach of McGraw and Barnes (2018). As time series with long memory or large noise levels are frequently encountered in the climate analyses that we are interested in, the ability to provide some formal estimate of model uncertainty is an important advantage of the approach adopted here.

3 Data and methods

We now describe an application of the above methods to sets of reanalysis-derived teleconnection indices. In addition to uncertainties due to model and parameter selection, in practice the reanalysis datasets that we use have associated uncertainties as well. Comparison of the networks derived from different products over a common timespan allows the role of these differences to be investigated and provides a baseline against which free-running model simulations can be compared.

3.1 Data

The 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).

The NCEP/NCAR Reanalysis 1 (Kalnay et al., 1996) 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 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; G. J. Marshall & Harangozo, 2000; G. J. 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 ($Z_{g500\text{ hPa}}$), zonal winds at 850 hPa and 200 hPa ($u_{850\text{ hPa}}$ and $u_{200\text{ hPa}}$), mean sea level pressure (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 $1^\circ{\times}1^\circ$ 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 correct issues found in previous reanalyses. 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 representation of the atmospheric circulation has been found to be greatly improved compared to previous generation reanalyses, although there remain known biases (Harada et al., 2016). Daily mean $Z_{g500\text{ hPa}}$, $u_{850\text{ hPa}}$, $u_{250\text{ hPa}}$, u_{sfc}, v_{sfc}, MSLP, and OLR fields are obtained on a $1.25^\circ{\times}1.25^\circ$ latitude-longitude grid. For SST fields, the model surface brightness temperature provided on a $1.25^\circ{\times}1.25^\circ$ latitude-longitude grid is used. Where required by the definition of the index as noted below, we regrid the initial fields to a $2.5^\circ \times 2.5^\circ$ latitude-longitude grid using a bilinear interpolation method.

3.2 Indices

From the full gridded fields we compute a set of indices diagnosing the activity of a selection of major global teleconnections, which will form the nodes in the fitted graphical models. In a fully data-driven approach, the definitions of such indices might be automatically determined by using community detection methods (Steinhaeuser et al., 2011; Bello et al., 2015). This has the advantage of accounting for differences between datasets or models in the representation of particular modes, e.g., due to shifts in the geographic centers of action. While these approaches have been employed in studies of causal effect networks (Kretschmer et al., 2017), here we use a set of fixed, empirical definitions for the teleconnection indices. We do so for two reasons. Firstly, this allows for a simpler evaluation of the performance of the models defined in Section 2, as the features in the fitted networks can be directly compared to well-studied relationships between traditional indices. Such a comparison would be more difficult to perform when using automatically extracted indices, as some differences might arise solely from the definition of the index. Additionally, positional shifts and other differences in the expression of particular modes with respect to a predefined diagnostic are themselves of interest from the point of view of comparing and evaluating models, and so we prefer to use a single set of reference definitions.

We choose a set of indices that provides reasonably comprehensive coverage of the dominant teleconnection processes active on intraseasonal through to interannual timescales. Where anomalies are required in the definition of an index, for consistency across the different datasets we compute all anomalies as differences from the daily or monthly climatology calculated with respect to the reference period 1 January 1979 to 30 December 2001.

As measures of tropical variability, we include an updated version (Zhang et al., 2019) of the multivariate ENSO index (MEI) (Wolter & Timlin, 1993, 1998, 2011), the dipole mode index (DMI) to characterize IOD activity (Saji et al., 1999), and the Wheeler-Hendon Madden-Julian oscillation (MJO) index (Wheeler & Hendon, 2004), denoted below as RMM1 and RMM2. 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. Where required, monthly MJO indices are defined as the monthly mean of the corresponding daily index.

In the extratropical atmosphere, we include indices of the AO, the SAM, the PNA, the PSA1 and PSA2 modes, and a set of modes associated with blocking in the North Atlantic and western Europe. We define the AO and SAM as the leading empirical orthogonal functions (EOFs) (Lorenz, 1956) of anomalies of monthly mean $Z_{g500\text{ hPa}}$ poleward of 20°N and 20°S, respectively. All anomalies are weighted by the square root of the cosine of the gridpoint latitude when computing the EOFs. Corresponding AO and SAM indices are calculated by projecting the (area-weighted) daily or monthly anomalies onto the leading EOF and normalizing by the standard deviation of the monthly leading principal component (PC).

The PNA pattern is taken to be the leading mode obtained after performing a VARIMAX rotation (Kaiser, 1958) of the first 10 EOF modes of monthly-standardized anomalies of monthly mean $Z_{g500\ \mathrm{hPa}}$ polewards of 20°N during boreal winter, taken to be 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 analogous modes in the SH, PSA1 and PSA2, are defined as the second and third modes in an EOF analysis of year-round anomalies of daily mean $Z_{g500\ \mathrm{hPa}}$ polewards of 20°S, projecting onto each mode and normalizing by the standard deviation of the correponding PC over the reference period to obtain associated indices.

Following the method presented in Straus et al. (2017), we define a set of Euro-Atlantic circulation regimes via a k-means clustering analysis of the leading 24 PCs of boreal winter anomalies in daily mean $Z_{g500\ \mathrm{hPa}}$ in the sector 20°N - 80°N, 90°W - 30°E, after applying a 10 day running mean smoothing. This method has the advantage of better capturing spatial asymmetries present in opposing phases of the NAO. Using $k = 4$ clusters, we obtain a set of patterns that correspond to the positive and negative NAO phases, NAO^+ and NAO^-, as well as two clusters associated with blocking events in the Atlantic and western Europe, denoted AR and SCAND, respectively. Indices for each of the 4 clusters are obtained by projecting daily or monthly height anomalies onto the composites associated with each cluster and standardizing by the monthly mean and standard deviations of the monthly index over the reference period.

For both reanalyses, monthly time series of the chosen indices are computed for the period 1 January 1960 to 30 November 2005, and this full time period is used for fitting DBNs. As some indices (e.g., SAM) exhibit significant trends over this period, we estimate and remove a linear trend for every index beforehand. Each index time series is then standardized to have zero mean and unit variance over the fitting period.

3.3 Priors and sampler settings

As noted in Section 2, when excluding the possibility of contemporaneous edges, it is natural to choose structurally modular priors for the parent sets $\mathrm{pa}_G(Y_t^i)$, such that the prior for a graph G decomposes into independent priors for the parent sets $\mathrm{pa}_G(Y_t^i)$ of each of the n indices included in the model. We fix a maximum allowed lag $\tau_{\max}$, such that the $n(\tau_{\max} + 1)$ nodes in the network at a given time are $\cup_{i=1}^n \{Y_t^i, \ldots, Y_{t-\tau_{\max}}^i\}$. For networks based on monthly indices, we take $\tau_{\max} = 6$ months, corresponding to the approximate e-folding time of the MEI. To enforce some degree of sparsity in the networks, we also impose a constraint on the maximum parent set size for each index, $|\mathrm{pa}_G(Y_t^i)| \leq p_{\max} = 10$. Subject to these constraints, in the absence of additional information we adopt uniform priors over the set of possible parent sets for each index, i.e.,

$$P(\mathrm{pa}_G(Y_t^i)) = \begin{cases} \left[\sum_{j=0}^{p_{\max}} \binom{n\tau_{\max}}{j}\right]^{-1}, & |\mathrm{pa}_G(Y_t^i)| \leq p_{\max}, \\ 0, & \text{otherwise}. \end{cases} \tag{16}$$

The MCMC sampling schemes described in Appendix A also require an appropriate proposal density for proposing updates to the structure of the model. We choose to 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 is defined as the set of graphs that can be reached from that structure by a single move in a predefined move set. Note that, when an explicit distinction is made between structure and parameter updates, as in the basic RJMCMC scheme in Appendix A, we do not include the current graph itself in the neighborhood. We take the set of possible moves to consist of 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 those graphs that can be reached from it by performing one of these three moves, subject to the imposed condition on the maximum parent set size. A structure update move thus consists of determining the neighborhood of the current graph based on the available moves before selecting with equal probability one graph from this set. As the models considered here allow for the node parameters to be marginalized out, we use the MC^3 sampling algorithm (Algorithm 2 in Appendix A) with the above proposal for fitting the model. For each index, posterior samples were obtained by running 8 chains of length 10×10^6 samples. The number of samples to be retained for analysis was chosen based on the estimated convergence rate from short initial runs following Brooks et al. (2003), where we required the thinning to be such that the resulting dependence between samples was reduced by a factor of 100 compared to the dependence between successive samples. To provide some assessment of chain convergence, homogeneity of the distribution of parent sets within chains was monitored using χ^2 and Kolmogorov-Smirnov tests (Brooks et al., 2003) for each index, although we note that, as always, these tests are not sufficient alone to determine approximate convergence (Sisson, 2005). In the following, both the posterior distribution over possible models and for individual features is of interest. Thus, in addition to applying these tests to the distribution of overall model indicators, we also monitored the results for the case when each model was labelled by a binary indicator for the presence of each possible edge, so as to check if inferences for the individual posterior edge probabilities were homogeneous across chains. Where required, samples for the node parameters were drawn from the conditional posterior distributions by Gibbs sampling.

The associations between the continuous valued indices are modeled using the linear Gaussian conditional density, Eq. (11). For each index, we take the hyperparameter values $a_\tau = 1.5$, $b_\tau = 20$, and $\nu_i^2 = 3$, for $i = 1, \dots, n$. This corresponds to independent t_3 marginal priors for the regression coefficients, with 95% prior HDI $-1 \leq \beta \leq 1$. The 1% and 99% percentiles for the conditional precision are respectively 1.1 and 113.4. Prior simulations suggest that this choice of priors yields a reasonable scale for the prior predictive distribution for the 1-step ahead forecast values, while not being so heavily regularized that relatively large values of the coefficients and indices are excluded. Similarly, the conditional precision hyperparameters are chosen to yield typical monthly innovation variances of order 1 or less. As these choices lead to somewhat informative priors, to assess the sensitivity of the results to the hyperparameter values we also perform fits with the 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), which we find lead to qualitatively similar results (see supporting information).

4 Results

4.1 Full year networks for monthly indices

We first consider the results of applying the above methods to year-round monthly indices for the two reanalyses. To better display the spatial structure of interactions, we separate the full network into subgraphs consisting of those indices corresponding to Northern Hemisphere (NH) extratropical, tropical, and SH extratropical modes, together with their inferred parent sets. From the posterior samples over possible parent sets for each index, approximate posterior probabilities $\hat{\pi}$ for the presence of each edge in the network are computed as the sample average, Eq. (4), of the corresponding indicator function.

The results are summarized in Figure 2 and Figure 3 for the possible parent sets of tropical indices, in Figure 4 for the NH extratropical indices, and in Figure 5 for the SH extratropical indices. For each subset, edges are shown with weights corresponding to their estimated posterior probability; for clarity, only those edges for which this value is at least 0.5 are shown. Overall, the learned structures for the two reanalyses are similar. In particular, there is broad agreement in terms of those edges for which the estimated posterior probability is high. These edges largely correspond, as is expected, to relationships in which there is a strong association between the parent and child node, either through strong autocorrelation or Granger causal links.

Strong evidence of long-range dependence on lagged values, reflecting the expected high levels of autocorrelation, is apparent for indices representing ENSO and the MJO, with the fitted models for both products featuring posterior probabilities near one for dependence of the MEI and both RMM indices on lags of up to four months (Figure 2 and Figure 3). In both cases, the preferred lags are consistent with the expected time-scales for these modes, noting that a maximum lag of 6 months is imposed when fitting the model. Purely atmospheric modes, on the other hand, exhibit little memory beyond time horizons of several weeks to a month. This is apparent in the fitted networks for both models in the absence of strong evidence for dependence of the NH extratropical (Figure 4) and SH extratropical indices (Figure 5) on lagged values of themselves beyond lags of one month. For instance, in both NNR1 and JRA-55 there is found to be little, or at most relatively weak, evidence for serial dependence on monthly time-scales for the AR and SCAND indices, which respectively represent blocking in the Atlantic and Scandinavia with a characteristic time-scale of 7 - 10 days. Although the DMI diagnoses Indian Ocean SST variability, the partial autocorrelation structure of the monthly mean DMI is consistent with an AR(1) process, and hence only the edge $\mathrm{DMI}_{t-1} \rightarrow \mathrm{DMI}_t$ is found to have appreciable posterior mass.

Where a high posterior probability edge is inferred between distinct indices, the edge generally matches with well-known associations or teleconnections among the modes included in the fit. This is highlighted, for example, in Figure 6, where the posterior probabilities for edges entering the AO, NAO^+, and NAO^- nodes in both reanalyses are shown. The close association between the AO and NAO is clearly evident in both reanalyses, and the estimated dependence relationships involving just these two indices are in essentially exact agreement. For both NNR1 and JRA-55, the presence of an edge in the

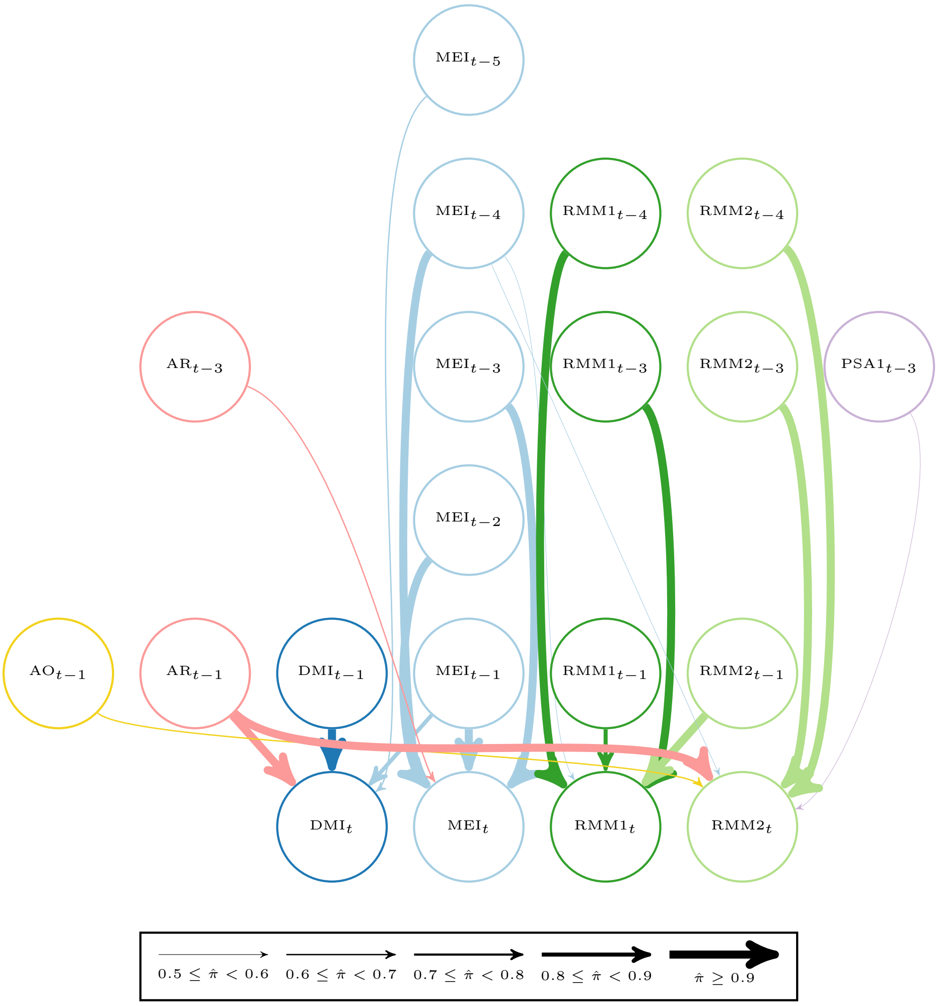


Figure 2. Subgraphs corresponding to the fitted parent sets of the tropical indices in NNR1 based on full-year data for $a_\tau = 1.5$, $b_\tau = 20$, and $\nu^2 = 3$. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.

network from the AO at lag 1 to each of the child nodes AO_t, NAO_t^+, and NAO_t^- is inferred with very high confidence. The same is true for the edges $\text{NAO}_{t-1}^- \rightarrow \text{AO}_t$ and $\text{NAO}_{t-1}^- \rightarrow \text{NAO}_t^+$. This clear interdependence of the AO and NAO is, of course, very well established, to the extent that the existence of the former as a distinct physical mode has been debated (Deser, 2000; Ambaum et al., 2001). The appearance of this relationship in the learned structures does, however, provide a useful check that the method recovers the expected relationships among particular modes. That these features agree between the two reanalyses also suggests that the relationships between these modes are consistent in the separate datasets.

Similarly, the fitted parent sets for the monthly PNA index in NNR1 and JRA-55, shown in Figure 7, both include with posterior probability greater than 0.5 dependence on the value of the PNA index at a lag of one month ($\hat{\pi} \approx 0.95$ and $\hat{\pi} \approx 0.91$ in NNR1 and JRA-55, respectively) and on the MEI at lag one ($\hat{\pi} \approx 0.71$ and $\hat{\pi} \approx 0.59$ for NNR1 and JRA-55, respectively). In this case, the results of the fit indicate that in both reanalyses there is reasonable evidence that the PNA is associated with tropical forcing, here captured by the lagged value of the MEI, consistent with previous observational and modeling studies (e.g., Hoskins & Ambrizzi, 1993; Trenberth et al., 1998; Franzke et al., 2011). The posterior mass for this edge is similar in NNR1 and JRA-55, suggesting that, as for the AO and NAO, this relationship is also consistent in the two products, in the sense that there is comparable evidence for the association across the two datasets.

Although the structures estimated from the NNR1 and JRA-55 datasets agree well in the above examples, there are notable differences in the estimated edge probabilities as well. Moreover, several of these differences involve features that are assigned appreciable posterior probability (i.e., $\hat{\pi} \geq 0.5$) based on the data from one reanalysis and not the other. That is, the differences are not limited only to associations that are weakly supported in both datasets. For example, in Figure 7, the estimated structure for NNR1 contains an edge from the lag one monthly mean RMM1 index to the monthly PNA index at lag 0 with high posterior probability. The same feature in JRA-55 is found to have a posterior probability $\hat{\pi} \approx 0.38$ that is less than 40% of that found for NNR1. Given the observed evidence for interactions between tropical convection and extratropical modes such as the PNA on intraseasonal time-scales (Lau & Phillips, 1986), in this case the weaker evidence for an MJO-PNA relationship in the JRA-55 data may arise due to the suppressed MJO observed in the JRA-55 reanalysis (Harada et al., 2016). It should be noted, however, that directly attributing the difference to this bias is not straightforward; for example, other known associations between the MJO and other modes of variability are identified in the JRA-55 fits, discussed below. Confirming whether there is a difference in this particular feature, and the underlying source of the difference, would require a closer examination of the representation of the two processes in the reanalysis, which is left for further studies. Nevertheless, this does highlight that comparison of the learned parent sets between the two reanalyses allows differences in the captured interactions between modes to be identified, particularly when we may have confidence that such a difference is in fact present. By adopting a Bayesian approach to structure learning, the level of confidence in this difference can be estimated: while the edge $\mathrm{RMM1}_{t-1} \to \mathrm{PNA}_t$ may be found in individual models for the time-evolution of the PNA index in both reanalyses, the existence of this feature is approximately 2.5 times more likely in NNR1 than in JRA-55 given the class of models and possible parent nodes that we consider. The presence of this edge with reasonable confidence in one product and not the other is suggestive of an underlying bias in one reanalysis, rather than simply being due to sampling variability.

Other differences in edges with large posterior mass, however, do not have quite as clear possible sources in specific underlying biases. In both reanalyses, the parent set of the NAO_t^+ node is estimated to contain the monthly PSA2 index at lag one, but with a posterior probability that is approximately 1.8 times larger in NNR1 ($\hat{\pi} \approx 0.89$ compared to $\hat{\pi} \approx 0.50$, see Figure 6). Similarly, the posterior probability for the edge $\mathrm{PSA2}_{t-1} \to \mathrm{NAO}_t^-$ is ≈ 4.7 times larger in NNR1, but in this case for both NNR1 and JRA-55 the estimated probability is less than 0.5, while an edge $\mathrm{PSA1}_{t-1} \to \mathrm{AR}_t$ is also present in both reanalyses. O'Kane et al. (2016) and O'Kane et al. (2017) concluded that the PSA modes predominantly reflect dynamics localized to within the SH waveguide, such that a direct physical interaction between the PSA2 mode and the NAO is unlikely. Thus, in contrast to the $\mathrm{RMM1}_{t-1} \to \mathrm{PNA}_t$ feature, one might expect that the differing level of confidence in this edge between the two reanalyses may be due to differences in relevant factors that are omitted from this simple analysis. In particular, as the fit shown in Figure 6 is based on data from all seasons, the presence of such a feature may reflect seasonal covariations in the extratropical circulation in both hemispheres that is otherwise unaccounted for here. While we consider the effects of seasonality in Section 4.2 below, ultimately direct determination of the root of these differences must be based on detailed evaluation of the two products. Here we seek only to highlight the use of the fitted networks for learning possible dependence relationships and hence their utility for guiding comparative analyses on the basis of identifying relationships that can be inferred to be present or absent with reasonable confidence.

Features for which there is lower confidence tend to differ more between the two reanalyses, although in these cases it becomes less clear as to whether they indicate substantive differences. This is most evident in the subgraphs corresponding to the parent sets for the SH extratropical indices in Figure 5. While there is agreement between the models fitted to NNR1 and JRA-55 in features such as the one month memory in the

SAM and the presence of an association (even if not a direct interaction) between the PSA1 and tropical forcing captured by the MEI, a larger number of features with posterior mass $\hat{\pi} > 0.5$ are found using the NNR1 data, involving a larger set of indices as parents. This greater disagreement between reanalyses, and overall lower confidence in the inferred non-independence relationships, may in part be due to the disparate resolutions and configurations of the respective atmospheric models in combination with relatively sparse observations in the SH prior to the satellite period. Spurious associations may also arise as a result of omitted factors, such as seasonal changes in the circulation, and low signal-to-noise ratios outside of the SH winter. For instance, the posterior distributions over parent nodes for the SAM index in the two reanalyses are summarized in Figure 8. In both NNR1 and JRA-55, non-zero associations are found with the MEI at lags of 4 and 6 months, albeit with somewhat higher posterior probabilities in JRA-55, and in NNR1 an edge from $\mathrm{AR}_{t-5} \to \mathrm{SAM}_t$ is identified with $\hat{\pi} \approx 0.52$. The same edge in JRA-55 is found to have $\hat{\pi} \approx 0.42$. Noting that a similar, comparatively low confidence relationship is found between NAO^+ and the PSA1 at a lag of four months in NNR1, and absent a mechanism for such interactions, it appears likely that in this case the association is an artifact arising from the use of year-round data; we confirm that this is the case in the following section. It is worth noting, however, that the fact that there is overall low posterior weight for these edges in both reanalyses allows identifying them as lacking robustness. This in turn is of use for the purposes of guiding model evaluation, where differences to observations in low probability relationships are of potentially less relevance.

It is important to bear in mind that the structures discussed above summarize the marginal posterior probability for the presence or absence of each individual edge over all possible models, rather than the presence or strength of an association between two indices within a single model. In addition to inspecting the estimated marginal distributions for each feature, given a sample from the approximate posterior distribution it may also be of interest to consider aspects of the sample that involve either the joint occurrence of one or more edges, as well as the posterior distribution over complete models. In particular, point-estimates for the parent set, analogous to those obtained using constraint-based approaches, can be obtained as the MAP estimate with the largest posterior probability. As all parent sets are assigned equal prior probability, this structure is simply the one that maximizes the marginal log-likelihood. Conditional on a particular MAP structure, estimates for the parameters under the model may be simply obtained by sampling from the conditional distributions $P(\theta|G, D)$, which in the case of the linear Gaussian model can be evaluated in closed form. Thus, in the sampling-based approach we may also obtain estimates of the strength of an association, conditioned on a particular model, in addition to the above estimates for the probability of the presence of the corresponding edge in the structure.

For example, the MAP parent sets for the NH extratropical modes in each reanalysis are summarized in Table 1, where we show both the estimated posterior probability for each edge in the parent set, as well as the posterior mean and 95% HDI for the corresponding coefficient in Eq. (11). In general, edges present in the MAP parent sets

Table 1. MAP parent sets for monthly NH extratropical teleconnection indices across all seasons for NNR1 and JRA-55 for fits with $a_\tau = 1.5$, $b_\tau = 20$, and $\nu^2 = 3$, showing the estimated posterior probability $\hat{\pi}$ of the edge, the mean parameter value $\hat{\beta}$ conditional on the MAP structure, and the 95% posterior HDI. Dashes indicate a node that is not in the MAP parent set for a given reanalysis.

	JRA-55			NNR1		
Parent node	$\hat{\pi}$	$\hat{\beta}$	95% HDI	$\hat{\pi}$	$\hat{\beta}$	95% HDI
AO_t						
AO_{t-1}	1.00	0.45	$(0.35, 0.56)$	1.00	0.46	$(0.35, 0.57)$
NAO^-_{t-1}	0.99	0.24	$(0.13, 0.35)$	0.99	0.25	$(0.13, 0.36)$
AR_t						
NAO^+_{t-6}	0.44	0.12	$(0.03, 0.19)$	0.32	0.12	$(0.04, 0.20)$
PNA_{t-1}	0.58	0.14	$(0.06, 0.22)$	0.36	0.14	$(0.06, 0.22)$
$\mathrm{PSA1}_{t-1}$	0.97	−0.17	$(-0.25, -0.09)$	0.67	−0.14	$(-0.22, -0.06)$
$\mathrm{RMM2}_{t-5}$	0.11	–	–	0.43	−0.12	$(-0.20, -0.04)$
NAO^+_t						
AO_{t-1}	1.00	0.35	$(0.23, 0.46)$	1.00	0.37	$(0.24, 0.48)$
NAO^+_{t-1}	0.66	0.17	$(0.05, 0.29)$	0.69	0.17	$(0.06, 0.29)$
NAO^-_{t-1}	0.98	0.31	$(0.19, 0.43)$	1.00	0.34	$(0.22, 0.46)$
$\mathrm{PSA2}_{t-1}$	0.50	−0.11	$(-0.19, -0.03)$	0.89	−0.14	$(-0.22, -0.06)$
NAO^-_t						
AO_{t-1}	1.00	−0.52	$(-0.65, -0.39)$	1.00	−0.52	$(-0.65, -0.39)$
NAO^-_{t-1}	0.78	−0.22	$(-0.34, -0.10)$	0.70	−0.23	$(-0.35, -0.11)$
PNA_{t-1}	0.73	−0.17	$(-0.26, -0.07)$	0.55	−0.15	$(-0.24, -0.06)$
$\mathrm{PSA2}_{t-1}$	0.10	–	–	0.47	0.12	$(0.04, 0.20)$
PNA_t						
MEI_{t-1}	0.58	0.21	$(0.12, 0.29)$	0.71	0.21	$(0.12, 0.29)$
PNA_{t-1}	0.91	0.15	$(0.07, 0.24)$	0.95	0.16	$(0.08, 0.24)$
$\mathrm{RMM1}_{t-1}$	0.38	–	–	0.99	0.17	$(0.09, 0.25)$
SCAND_t						
MEI_{t-6}	0.34	−0.13	$(-0.21, -0.05)$	0.28	–	–
SCAND_{t-1}	0.53	0.13	$(0.04, 0.21)$	0.31	–	–
MEI_{t-5}	0.23	–	–	0.45	−0.13	$(-0.22, -0.05)$
AR_{t-1}	0.28	–	–	0.53	−0.13	$(-0.21, -0.05)$

tend to have at least moderately high posterior probabilities, that is, they correspond to non-independence relationships that are also present in a large fraction of the sampled models. Edges found in the MAP parent set are not necessarily found in the majority of possible models, however, with several edges having posterior probabilities of order ~ 0.3 in Table 1. For these edges, while their inclusion leads to a good fit for this particular sample, the data do not provide especially strong evidence for their presence compared to other edges in the MAP structure, given the set of all other possible predictors. To some extent this also reflects the fact that multiple models with parent sets that do not include these edges may still provide a reasonable fit to the observed data, despite not maximizing the marginal likelihood, and hence can account for non-negligible posterior mass. Considering only the single most probable structure may therefore fail to take into account relevant model uncertainty. Notably, differences between the MAP structures for the two reanalyses are again not only restricted to edges with low posterior mass, suggesting that at least some differences may be due to systematic effects rather than as a result of minor differences in the two samples. For example, the absence of the edge $\mathrm{RMM}_{t-1} \to \mathrm{PNA}_t$ in the MAP parent set from JRA-55 indicates that not only is there weak evidence for this relationship in the dataset, but that it is also not required in order to produce a good fit to the observed time series. Where an edge is present in the MAP parent set for both reanalyses, there is good agreement between the two products in terms of the estimated coefficient. For the NH modes, the two reanalyses yield very similar estimates for the strength of each association, while there are somewhat larger differences for the tropical and SH extratropical modes (see supporting information).

4.2 Seasonal networks for monthly indices

Overall, the fitted networks based on full year data show good agreement between the two reanalyses. Some of the differences noted in the previous section may arise due to confounding or spurious associations generated by, among other things, the omission of relevant variables in the fit (i.e., failure of causal sufficiency, in the case that the models are interpreted as being causal). An obvious possible factor is the seasonal variation in relationships between nodes that arises as a result of seasonal changes in the background flow and hence in the available pathways for propagation of disturbances (e.g., Hoskins & Ambrizzi, 1993; Ambrizzi et al., 1995). To account for this seasonal dependence, one possibility would be to include a seasonal indicator as a node within the graph itself. For simplicity, however, to investigate the impact of seasonal variation we consider the results of repeating the above analysis restricted to data in the three month winter season for each hemisphere. Note that, as lags of up to six months are still allowed, observations entering into these fits also include lagged values of the indices during the previous season. For brevity, we restrict our attention to only a subset of the major teleconnections within each hemisphere.

In Figure 9, the estimated posterior probabilities for the parent sets of the NAO^+ and NAO^- indices during DJF are shown. As for the full year fits, in both reanalyses there is strong evidence for an association between the AO and the two phases of the NAO, with approximately equal posterior mass assigned to each edge in NNR1 and JRA-55. Compared to the full year networks, the edge from $\text{PSA2}_{t-1} \rightarrow \text{NAO}_t^+$ is no longer found to have appreciable posterior mass ($\hat{\pi} < 0.1$ in both reanalyses), consistent with this edge arising as a result of the use of full-year data. In both NNR1 and JRA-55, a relationship between the MJO, via the value of the RMM1 index at a lag of two months, with the positive phase of the NAO is found in a large fraction of sampled models ($\hat{\pi} \approx 0.67$ in NNR1 and $\hat{\pi} \approx 0.83$ in JRA-55). Interactions between the MJO and the NAO have previously been reported in winter season observations (e.g., Lin et al., 2009) arising from known dynamical mechanisms (e.g., Frederiksen & Frederiksen, 1993). The inferred presence of this relation, with moderately high confidence, indicates that both products are consistent in capturing this association, although it should be noted that monthly mean data is being used here in contrast to the more usual daily or pentad data. In both NNR1 and JRA-55, additional edges are also found with somewhat lower posterior probabilities, including the edges $\text{NAO}_{t-5}^+ \rightarrow \text{NAO}_t^-$, $\text{PSA1}_{t-5} \rightarrow \text{NAO}_t^-$, and $\text{PNA}_{t-3} \rightarrow \text{NAO}_t^+$ with estimated probabilities $\hat{\pi} \approx 0.55, 0.51$, and 0.54 in JRA-55, respectively, and $\hat{\pi} \approx 0.49, 0.43$, and 0.45 in NNR1. As these features are present in both reanalyses with relatively similar posterior weights, this suggests that there is some, if comparatively weak, evidence for these associations from both products, and the two reanalyses appear to thus be consistent. More notably, the feature $\text{DMI}_{t-6} \rightarrow \text{NAO}_t^-$ is estimated to have a posterior probability of $\hat{\pi} \approx 0.52$ based on the indices computed using NNR1 and HadISST data, while when fitted to JRA-55 the same feature is assigned a posterior mass of only $\hat{\pi} \approx 0.09$, implying substantially weaker evidence is found for the presence of this edge in the JRA-55 data.

For fits based on austral winter (June-July-August, JJA) data, similar results are found in that several apparently spurious associations cease to be present in a large fraction of the sampled structures. In Figure 10, the estimated posterior probabilities for parent nodes of the PSA1 index during JJA are shown for the two reanalyses. Compared to the previous, full-year fits in Figure 5, the set of edges with high posterior mass from NNR1 data no longer includes long-range dependence on the positive phase of the NAO, or on the PSA2 index. Instead, an edge $\text{RMM1}_{t-5} \to \text{PSA1}_t$ is found with posterior probability $\hat{\pi} \approx 0.60$, with the same feature being obtained from the JRA-55 data with $\hat{\pi} \approx 0.77$. The JRA-55 fit also contains an edge $\text{RMM2}_{t-6} \to \text{PSA1}_t$ with estimated posterior probability $\hat{\pi} \approx 0.94$; in NNR1, the corresponding feature is found to have $\hat{\pi} \approx 0.42$. Associations between the MJO and the PSA modes on intraseasonal time-scales during winter have previously been noted (Mo & Paegle, 2001), although the fitted models here assign greater posterior weight to dependence at longer lags. While the posterior probability for the presence of RMM1 as a predictor of the monthly mean PSA1 is roughly consistent between NNR1 and JRA-55, the approximate factor of two difference in the value of the sampled posterior probability for the RMM2 edge is more sizable and may provide weak evidence of a difference between the two reanalyses in terms of this relationship. A similar statement may be made for the $\text{MEI}_{t-1} \to \text{PSA1}_t$ edge, for which the fitted posterior probabilities are $\hat{\pi} \approx 0.59$ in JRA-55 and $\hat{\pi} \approx 0.22$ in NNR1. Interestingly, approximately half ($\hat{\pi} \approx 0.49$) of the sampled structures in NNR1 instead contain an edge $\text{PNA}_{t-1} \to \text{PSA1}_t$, which may reflect the effects of a common dependence on tropical forcing. Overall, there are a larger number of features with high posterior mass in the JRA-55 fits. As usual, the precise underlying reasons for these differences are not determined by the fits alone. Possible biases, such as a poorer representation of the wintertime SH circulation in one product compared to the other, would require further detailed follow-up. Our purpose here has been to highlight the use of Bayesian structure learning as a tool to identify possible differences, and to estimate the level of uncertainty associated with each.

5 Summary

Probabilistic graphical models provide a natural and intuitive framework with which to describe the complicated interactions between climate processes. As a result, they are increasingly being applied for the purposes of studying potential causal relationships and for model evaluation (Vázquez-Patiño et al., 2020; Nowack et al., 2020). In this latter application, models are generally evaluated on the basis of structural comparisons between graph structures inferred from observations and from model runs. Additionally, the strength of associations may be compared by performing a second stage of model fitting, conditional on the inferred structure. As is, this constitutes a powerful lens for examining differences between models and observations.

That being said, existing approaches still have some important limitations when used as tools for model evaluation. In particular, the most widely used strategy of first learning a suitable structure using a constraint-based learning algorithm, followed optionally by fitting a regression model conditioned on this structure, does not lend itself to easily estimating the level of confidence in the obtained model. While sensitivity analyses enable some determination of the robustness of particular features, in general the sampling properties associated with this procedure are difficult to assess (Madigan & Raftery, 1994; Draper, 1995). This can present a challenge for using learned structures as tools for model evaluation, where some measure of significance of observed differences is usually desirable so as to assess whether they are due to model biases or sampling variability.

This limitation can in principle be overcome by employing a Bayesian approach to structure learning. By learning a posterior distribution over possible structures, rather than selecting a single graph, overall model uncertainty can be quantified and accounted for. This can be particularly important where multiple different structures may all be nearly equally well supported by the data, in which case selection of a single model may overestimate the confidence warranted in particular features. Given a sample from the model posterior distribution, by averaging over the set of possible models the posterior credibility of given features may instead be estimated in the Bayesian approach to identify edges that are well supported by the data. Subsequent estimation of the model parameters conditional on a given structure is straightforward, and provides a basis for comparing the magnitude of relationships between different processes.

The result of the sampling-based structure learning algorithms is a sample from the set of possible models, from which posterior probabilities for particular features can be derived. In this way, robust features for which there is high confidence may be identified, and the set of such edges may in turn form the focus of model comparisons. To illustrate this approach, we have applied an MCMC based approach to learn DBNs describing associations between teleconnections in two different reanalyses, with the goals of identifying the robust structural differences between the two products and to establish a set of baseline estimates for subsequent model evaluation studies.

In general, features in the networks derived from NNR1 and JRA-55 data that have high estimated posterior probabilities agree reasonably well. While not surprising, as both reanalyses attempt to provide an estimate of the same climate state, this provides some reassurance that consistent results are obtained using a sampling algorithm, and that the expected associations and dependence structures, such as long memory in oceanic modes and close correspondence between the AO and NAO, are recovered with reasonable levels of certainty. Differences between the models estimated from the two reanalyses are not only limited to edges with low posterior mass, however. In some cases, these differences involve modes for which there are known biases in one reanalysis or the other; the lack of evidence for a dependence of the PNA on the MJO in JRA-55 is one such example. In other cases, these features may arise as a result of the effects of omitted confounding effects, such as seasonal cycles and other common drivers, that nevertheless differ between the two reanalyses. Some evidence for this is found by considering networks fitted from data restricted to only the winter season in each hemisphere. In these fits, apparently spurious cross-equatorial dependence present in fits to year-round data are no longer found to have strong evidence to support their presence. A greater number of differences between the networks derived from the two reanalyses are found for edges that have lower posterior probabilities, with this being particularly noticeable for the SH modes. This may in part be due to greater differences in the representation of these modes, as well as lower signal-to-noise ratios, at least outside of the austral winter.

It is important to note that in this study our aim has not been to perform a detailed evaluation of the differences between the two reanalysis products. Extensive characterization of the performance and biases of both NNR1 and JRA-55 has been done in the past, and further investigations of the differences found here would require additional detailed study of the involved processes in each product. Rather, our purpose has been to demonstrate the applicability of the Bayesian approach to structure learning in deriving graphical models suitable for use as tools for process-based model evaluation. We argue that an important aspect of this application is the need to account for inevitable model uncertainty in order to identify differences that are likely to be robust and hence represent genuine model biases. Independent of the context of the analysis, this can be naturally achieved in the Bayesian approach. By considering a relatively straightforward initial application to reanalysis data, for which there exists (at least some) consensus on the interactions between modes, we have shown that a score-based sampling approach recovers the expected relationships, while also providing additional benefits over constraint-based approaches in the form of estimates for the posterior distribution over models and features.

Given this, we have primarily focused on a single analysis of year-round monthly mean data, or data within a single season. While sufficient to illustrate the relevant features of the results, in order to utilize this approach in the context of causal discovery it would be necessary to further extend the analysis presented here. In addition to careful selection of the relevant variables, time periods, and temporal resolution, further consideration should be given to the form of the likelihood and priors used in defining the model. In this work, the simplest case of a linear model with conjugate priors on the parameters defining the conditional PDFs has been used, 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. More complex forms for the conditional PDF, as well as the use of non-conjugate priors and inclusion of additional constraints, may be directly handled using a generic reversible jump MCMC instead of the more specialized MC^3 used here.

Whether or not the resulting model has analytic structure, care must be taken in the design of the sampler and in assessing the approximate convergence of the simulation to the target posterior distribution. For the results presented here, the non-parametric convergence diagnostics of Brooks et al. (2003) have been used to monitor (non-)convergence, but in general assessing convergence for trans-dimensional MCMC remains challenging (Sisson, 2005). In our case, trace plots and convergence diagnostics applied to individual edge indicators suggested that (model-averaged) estimates derived from individual chains were consistent and stable, but there was evidence for non-homogeneity in the posterior distribution over structures across chains based on 10×10^6 samples. For this reason, we have avoided making definitive statements with respect to Bayes factors and other model-specific quantities, recognizing that further sampling would be required for these to be reliably determined. Poor mixing and multi-modal posterior distributions over models may also be problematic, with the sampled chains remaining trapped in the vicinity of a single mode. Given the large model space considered here, this is a concern as ensuring the chains have adequately explored the full posterior distribution is unlikely to be feasible, raising the possibility that the fits have converged to a local mode in the model space. For generating causal hypotheses, it may be more suitable to restrict attention to smaller systems of variables. On the other hand, the relatively close agreement between the fits for the two reanalyses shown here suggests that the simulations explore sufficiently similar regions of the model space for useful qualitative comparisons to still be performed. As in this paper only two datasets were compared, direct inspection of the individual fitted networks, amounting to visual inspection of the sampled posterior distributions, was sufficient. In more extensive evaluation studies, this may be extended by the use of appropriate summary measures computed on the fits for different models.

The DBN models obtained for the NNR1 and JRA-55 reanalyses form a set of ground truth results against which free-running models may be compared; we are currently performing such a comparison over the historical period. However, the homogeneous models considered here cannot model one of the most notable features of the observed climate over this period, namely, the existence of secular trends in the behavior of particular modes. Key open questions remain as to whether there is evidence for accompanying changes in the underlying interaction structures, and whether models indicate the possibility of regime shifts under future forcing scenarios. Addressing these questions in this framework will require the use of non-homogeneous network models in which either the model parameters or structure are allowed to vary. In a forthcoming article, we perform just such a comparison against the reanalysis-derived networks presented here for a subset of the CMIP5 ensemble, as well as relaxing the time-homogeneity assumptions discussed above to study the existence of regime transitions in projections. Assessing the evidence for, say, the existence of sudden structure changes versus slow variation in the underlying network parameters does not entail any additional conceptual changes, making the approach presented here well-suited for investigating such questions.

Acknowledgments

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).

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 and the generated indices used to perform the analyses presented in this study may be found at `https://doi.org/10.5281/zenodo.4331149`.

DH is supported by the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO) through a ResearchPlus postdoctoral fellowship. TO is supported by the CSIRO Decadal Climate Forecasting Project (`https://research.csiro.au/dfp`).

Appendix A MCMC sampling methods

In this appendix we summarize the MCMC methods used for structure learning. We employ the composite parameter space formulation described in detail in Godsill (2001). For the models considered in the main text, in the absence of same-time edges the networks are structurally modular and the parent set associated with each index can be inferred independently. However, in general a given structure $G \in \mathcal{G}$ will describe the joint distribution of all of the indices simultaneously at a given time. To accommodate such models, we keep the notation relatively general.

The observed data takes the form of a time series of one or more indices $D = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$. The allowed set of models $\mathcal{G}$ to describe these data are taken to be specified a priori. Let $\boldsymbol{\theta}$ denote the collection of all such parameters across all of the possible models in $\mathcal{G}$. For example, for the set of all linear Gaussian models, the vector $\boldsymbol{\theta}$ would contain all of the possible regression coefficients $\beta^i_{(j,\tau)}$ associated with the possible edges $Y^j_{t-\tau} \to Y^i_t$ as well as the conditional precisions τ^2_i and intercept parameters β^i_0. Any given model G

will only use a small subset of all of the possible parameters. The subset of the parameters used by a structure G will be written $\boldsymbol{\theta}_{\mathcal{I}(G)}$; here $\mathcal{I}(G)$ is an appropriate index set indicating which components of the complete parameter vector $\boldsymbol{\theta}$ are required by G. The remaining parameters not used by G will be denoted by $\boldsymbol{\theta}_{-\mathcal{I}(G)}$. A particular model for the indices $\{Y^1, \dots, Y^n\}$ is described by the pair $(G, \boldsymbol{\theta}) \in \mathcal{G} \times \Theta$, where $\mathcal{G} \times \Theta$ is the composite model space (Godsill, 2001).

To infer the network structure and parameters given data D, we aim to sample from the posterior distribution

$$P(G, \boldsymbol{\theta}|D) \propto P(D|G, \boldsymbol{\theta})P(G, \boldsymbol{\theta}).$$

The likelihood $P(D|G, \boldsymbol{\theta})$ is taken to depend only on the subset of parameters associated with G,

$$P(D|G, \boldsymbol{\theta}) \equiv P(D|G, \boldsymbol{\theta}_{\mathcal{I}(G)}),$$

while the prior distribution $P(G, \boldsymbol{\theta})$ is taken to be of the form

$$P(G, \boldsymbol{\theta}) = P(\boldsymbol{\theta}_{\mathcal{I}(G)}|G)P(\boldsymbol{\theta}_{-\mathcal{I}(G)}|\boldsymbol{\theta}_{\mathcal{I}(G)}, G)P(G).$$

The factor $P(\boldsymbol{\theta}_{-\mathcal{I}(G)}|\boldsymbol{\theta}_{\mathcal{I}(G)}, G)$ corresponds to a set of proper pseudo-priors (Carlin & Chib, 1995) for the parameters not used by G, and may otherwise be chosen essentially freely. The samplers that we use correspond to Metropolis-Hastings schemes in the composite model space (Godsill, 2001) with a proposal density of the form

$$q(G', \boldsymbol{\theta}'; G, \boldsymbol{\theta}) = q_1(G'; G)q_2(\boldsymbol{\theta}'_{\mathcal{I}(G')}; \boldsymbol{\theta}_{\mathcal{I}(G)})P(\boldsymbol{\theta}'_{-\mathcal{I}(G')}|\boldsymbol{\theta}'_{\mathcal{I}(G')}, G'), \tag{A1}$$

for a move from $(G, \boldsymbol{\theta})$ to $(G', \boldsymbol{\theta}')$ with corresponding acceptance probability

$$\alpha = \min\left\{1, \frac{q_1(G; G')}{q_1(G'; G)} \frac{q_2(\boldsymbol{\theta}_{\mathcal{I}(G)}; \boldsymbol{\theta}'_{\mathcal{I}(G')})}{q_2(\boldsymbol{\theta}'_{\mathcal{I}(G')}; \boldsymbol{\theta}_{\mathcal{I}(G)})} \frac{P(G', \boldsymbol{\theta}'_{\mathcal{I}(G')}|D)}{P(G, \boldsymbol{\theta}_{\mathcal{I}(G)}|D)}\right\}. \tag{A2}$$

When the class of models considered does not admit analytic evaluation of any of the required integrals, we make use of the simple reversible jump MCMC scheme given in Algorithm 1. At each iteration, either an update to the parameter associated with the structure G is chosen, with probability $j_\theta(G, \boldsymbol{\theta}_{\mathcal{I}(G)})$, or an update to the current structure is proposed. In the case of a parameter update, 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 updated state $(G, \boldsymbol{\theta}'_{\mathcal{I}(G)})$ is accepted with probability

$$\alpha = \min\left\{1, \frac{j_\theta(G, \boldsymbol{\theta}'_{\mathcal{I}(G)})}{j_\theta(G, \boldsymbol{\theta}_{\mathcal{I}(G)})} \frac{q_\theta(\boldsymbol{\theta}_{\mathcal{I}(G)}; \boldsymbol{\theta}'_{\mathcal{I}(G)})}{q_\theta(\boldsymbol{\theta}'_{\mathcal{I}(G)}; \boldsymbol{\theta}_{\mathcal{I}(G)})} \frac{P(G, \boldsymbol{\theta}'_{\mathcal{I}(G)}|D)}{P(G, \boldsymbol{\theta}_{\mathcal{I}(G)}|D)}\right\}. \tag{A3}$$

Note that, when the probability of a parameter update move is the same in both states, this is simply an ordinary Metropolis-Hastings update for a single model. If instead a structure update move is chosen, a new structure G' is proposed according to the proposal distribution $q_G(G'; G)$. Any parameters that are common to both G' and G are held fixed at their current values, while any new parameters $\boldsymbol{\theta}'_{\mathcal{I}(G')\setminus\mathcal{I}(G)}$ are sampled from an additional proposal density $\tilde{q}_\theta(\boldsymbol{\theta}'_{\mathcal{I}(G')\setminus\mathcal{I}(G)})$; if $\mathcal{I}(G') \subset \mathcal{I}(G)$ we set $\tilde{q}_\theta \to 1$. Parameters that are either present in the initial structure but not in the proposed structure, or are not used by either, are left unchanged for simplicity. The acceptance ratio for this move is

$$\alpha = \min\left\{1, \frac{j_G(G', \boldsymbol{\theta}'_{\mathcal{I}(G')})}{j_G(G, \boldsymbol{\theta}_{\mathcal{I}(G)})} \frac{q_G(G; G')}{q_G(G'; G)} \frac{\tilde{q}_\theta(\boldsymbol{\theta}_{\mathcal{I}(G)\setminus\mathcal{I}(G')})}{\tilde{q}_\theta(\boldsymbol{\theta}'_{\mathcal{I}(G')\setminus\mathcal{I}(G)})} \frac{P(G', \boldsymbol{\theta}'_{\mathcal{I}(G')}|D)}{P(G, \boldsymbol{\theta}_{\mathcal{I}(G)}|D)}\right\}. \tag{A4}$$

This structure update move is just a particular case of the general reversible jump move (Green, 1995) with a unit Jacobian. For more general mappings from the current to proposed parameters a non-trivial Jacobian factor would remain following the change of variables in the proposal density Eq. (A1).

When the models considered allow for the conditional posterior distribution $P(\boldsymbol{\theta}_{\mathcal{A}(G)}|\boldsymbol{\theta}_{\mathcal{I}(G)\setminus\mathcal{A}(G)}, G, D)$ of some subset of the parameters $\boldsymbol{\theta}_{\mathcal{A}(G)}$, $\mathcal{A}(G) \subseteq \mathcal{I}(G)$, to be evaluated given G and any remaining parameters $\boldsymbol{\theta}_{\mathcal{I}(G)\setminus\mathcal{A}(G)}$, we adopt a conditional Metropolis-Hastings scheme that takes better advantage of this structure. For simplicity, we assume that the set of parameters $\boldsymbol{\theta}_{\mathcal{I}(G)\setminus\mathcal{A}(G)}$, if not empty, is shared across all of the possible structures. Updates to the structure are proposed as before under the proposal $q_G(G'; G)$. The shared parameters $\boldsymbol{\theta}_{\mathcal{I}(G)\setminus\mathcal{A}(G)}$ are kept at their previous values, while the remaining parameters are drawn from the exactly known posterior distribution $P(\boldsymbol{\theta}'_{\mathcal{A}(G')}|\boldsymbol{\theta}'_{\mathcal{I}(G')\setminus\mathcal{A}(G')}, G', D)$. Under this proposal, the acceptance ratio simplifies to

$$\alpha = \min\left\{1, \frac{q_G(G; G')}{q_G(G'; G)} \frac{P(G'|\boldsymbol{\theta}'_{\mathcal{I}(G')\setminus\mathcal{A}(G')}, D)}{P(G|\boldsymbol{\theta}_{\mathcal{I}(G)\setminus\mathcal{A}(G)}, D)}\right\}. \tag{A5}$$

The parameters associated with the current structure can be updated via a standard Metropolis-Hastings or Gibbs step, as in the parameter update move for the simple reversible jump scheme. Parameter and structure updates can either be proposed randomly or performed in a fixed order. For the models presented in Section 4, the conditional posterior distribution for all of the parameters can be evaluated, i.e., the set $\boldsymbol{\theta}_{\mathcal{I}(G)\setminus\mathcal{A}(G)}$ is empty. In this case, the scheme reduces to the MC^3 scheme of Madigan et al. (1995), with the dependence on the graph parameters dropping out entirely. The acceptance ratio for a struc-

Algorithm 1 Simple RJMCMC sampler

Require: initial state $x^{(1)} \equiv (G^{(1)}, \boldsymbol{\theta}^{(1)})$, observed data D, chain length $2S$

for $s = 2, \ldots, 2S$ **do**
  Draw $u_1, u_2 \sim \text{Uniform}(0,1)$
  if $u_1 < j_\theta(x^{(s-1)})$ **then**
    Set $G' \leftarrow G^{(s-1)}$
    Draw new parameters $\boldsymbol{\theta}'_{\mathcal{I}(G')}$ from $q_\theta(\boldsymbol{\theta}'_{\mathcal{I}(G')}; \boldsymbol{\theta}^{(s-1)}_{\mathcal{I}(G^{(s-1)})})$
    Calculate α according to Eq. (A3)
  else
    Draw new parent set G' from $q_G(G'; G^{(s-1)})$
    Draw $\boldsymbol{\theta}'_{\mathcal{I}(G') \setminus \mathcal{I}(G^{(s-1)})}$ from $\tilde{q}_\theta(\boldsymbol{\theta}'_{\mathcal{I}(G') \setminus \mathcal{I}(G^{(s-1)})})$
    Calculate α according to Eq. (A4)
  if $u_2 < \alpha$ **then**
    $x^{(s)} \leftarrow (G', \boldsymbol{\theta}')$
  else
    $x^{(s)} \leftarrow (G^{(s-1)}, \boldsymbol{\theta}^{(s-1)})$
Discard first S samples as warm-up
return $\{x^{(S+1)}, \ldots, x^{(2S)}\}$

ture drawn according to $q_G(G'; G)$ is in this case

$$\alpha = \min\left\{1, \frac{q_G(G; G')}{q_G(G'; G)} \frac{P(D|G')}{P(D|G)} \frac{P(G')}{P(G)}\right\}, \tag{A6}$$

where the likelihood can be written in terms of the local marginal likelihoods, as in Eq. (8). Particular choices of the proposal density and structure priors are described in Section 3.3. We summarize the resulting sampling scheme for the structures in Algorithm 2. For each sampling method, we run multiple chains, discarding the first half of each sample as burn-in. For the MC^3 sampler, approximate convergence of the chains to the target distribution is assessed using the χ^2 and Kolmogorov-Smirnov tests proposed in Brooks et al. (2003).

Algorithm 2 MC^3 sampler

Require: initial state $G^{(1)}$, observed data D, chain length $2S$

```
for s = 2, ..., 2S do
    Draw u ~ Uniform(0, 1)
    Draw new parent set G' from q_G(G'; G^(s-1))
    Calculate α according to Eq. (A6)
    if u < α then
        G^(s) ← G'
    else
        G^(s) ← G^(s-1)
Discard first S samples as warm-up
return {G^(S+1), ..., G^(2S)}
```

Appendix B Expressions for marginal likelihoods

In this appendix we state the 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 in the main text.

The linear Gaussian model given in Section 2 reads

$$
\begin{aligned}
\tau_i^2 &\sim \mathrm{Gamma}(a_\tau, b_\tau),\\
\beta_0^i|\tau_i^2, \nu_i^2 &\sim N\left(0, \frac{\nu_i^2}{\tau_i^2}\right),\\
\beta_{(k_j,\tau_j)}^i|\tau_i^2, \nu_i^2, \mathrm{pa}_G(Y_t^i) &\sim N\left(0, \frac{\nu_i^2}{\tau_i^2}\right), \quad j = 1, \dots, |\mathrm{pa}_G(Y_t^i)|,\\
Y_t^i|\beta_0^i, \beta_{(k_j,\tau_j)}^i, \tau_i^2, \mathrm{pa}_G(Y_T^i) &\sim N\left(\beta_0^i + \sum_{j=1}^{|\mathrm{pa}_G(Y_t^i)|} \beta_{(k_j,\tau_j)}^i Y_{t-\tau_j}^{k_j}, \frac{1}{\tau_i^2}\right),
\end{aligned}
$$

For the prior on the conditional precision τ_i^2, we adopt the convention

$$
P(\tau_i^2|a_\tau, b_\tau) = \frac{1}{\Gamma(a_\tau)} \frac{(\tau_i^2)^{a_\tau - 1}}{b_\tau^{a_\tau}} \exp\left(-\frac{\tau_i^2}{b_\tau}\right).
$$

With this convention, the unconditional prior distribution for a given coefficient β is a generalized t-distribution with scale $\hat{\sigma}^2 = \nu^2/(a_\tau b_\tau)$ and $2a_\tau$ degrees of freedom, $\beta \sim T_1(0, \nu^2/(a_\tau b_\tau), 2a_\tau)$. For the models shown in Section 4, a_τ, b_τ, and ν_i^2 are taken as fixed hyperparameters. In practice, the signal-to-noise ν_i^2 may be poorly known, in which case it is possible to also sample it from a conjugate inverse gamma prior using the sampling schemes for models with partial analytic structure discussed in Appendix A. For a given parent set, we write the vector of predictor variables as ($p_i = |\mathrm{pa}_G(Y_t^i)|$)

$$\boldsymbol{x}_t^{iT} = (1, y_{t-\tau_1}^{k_1}, \ldots, y_{t-\tau_{p_i}}^{k_{p_i}}),$$

at each time $t = 1, \ldots, T$ and introduce the $T \times (p_i + 1)$ design matrix

$$X_i = \begin{pmatrix} \boldsymbol{x}_1^{iT} \\ \vdots \\ \boldsymbol{x}_T^{iT} \end{pmatrix}.$$

The likelihood for the observed values of the index Y^i then takes the simple form of a product of normal densities, and the local marginal likelihood $\Psi_i(D; G)$ can be evaluated using standard conjugacy results. The marginal joint distribution for the observed index values $\boldsymbol{y}_i^T = (y_1^i, \ldots, y_T^i)$ under this model is a multivariate t-distribution, giving

$$\Psi_i(D; G) = \frac{\Gamma\left(\frac{T+2a_\tau}{2}\right)}{\Gamma\left(\frac{2a_\tau}{2}\right)\pi^{T/2}(2a_\tau)^{T/2}} (\det \Sigma_i)^{-1/2} \left(1 + \frac{1}{2a_\tau}\boldsymbol{y}_i^T \Sigma_i^{-1} \boldsymbol{y}_i\right)^{-\frac{T+2a_\tau}{2}}, \tag{B1}$$

where

$$\Sigma_i = \frac{1}{a_\tau b_\tau}\left(I_{T\times T} + \nu_i^2 X_i X_i^T\right), \tag{B2}$$

and $I_{T\times T}$ is the $T\times T$ identity matrix. For a given parent set, the regression coefficients and precision can be sampled from the conditional posterior distributions

$$\tau_i^2 | D, G, \nu_i^2 \sim \mathrm{Gamma}\left(\frac{T + 2a_\tau}{2}, \frac{b_\tau}{1 + \frac{1}{2a_\tau}\boldsymbol{y}_i^T \Sigma_i^{-1} \boldsymbol{y}_i}\right), \tag{B3}$$

$$\boldsymbol{\beta}_i | \tau_i^2, \nu_i^2, D, G \sim N\left(\tau_i^2 \Sigma_{\beta_i} X_i^T \boldsymbol{y}_i, \Sigma_{\beta_i}\right), \tag{B4}$$

where the posterior covariance matrix for the coefficients $\boldsymbol{\beta}_i^T \equiv (\beta_0^i, \beta_{(k_1,\tau_1)}^i, \ldots, \beta_{(k_{p_i},\tau_{p_i})}^i)$ is given by

$$\Sigma_{\beta_i} = \frac{\nu_i^2}{\tau_i^2}\left(I_{(p_i+1)\times(p_i+1)} + \nu_i^2 X_i^T X_i\right)^{-1}.$$

References

Abramson, B., Brown, J., Edwards, W., Murphy, A., & Winkler, R. L. (1996). Hailfinder: A Bayesian system for forecasting severe weather. *International Journal of Forecasting*, *12*(1), 57 - 71. doi: 10.1016/0169-2070(95)00664-8

Ambaum, M. H. P., Hoskins, B. J., & Stephenson, D. B. (2001). Arctic Oscillation or North Atlantic Oscillation? *Journal of Climate*, *14*(16), 3495-3507. doi: 10.1175/1520-0442(2001)014⟨3495:AOONAO⟩2.0.CO;2

Ambrizzi, T., Hoskins, B. J., & Hsu, H.-H. (1995). Rossby Wave Propagation and Teleconnection Patterns in the Austral Winter. *Journal of the Atmospheric Sciences*, *52*(21), 3661-3672. doi: 10.1175/1520-0469(1995)052⟨3661:RWPATP⟩2.0.CO;2

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

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

Berezin, Y., Gozolchiani, A., Guez, O., & Havlin, S. (2012). Stability of Climate Networks with Time. *Scientific Reports*, *2*(1), 666. doi: 10.1038/srep00666

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

Boneh, T., Weymouth, G. T., Newham, P., Potts, R., Bally, J., Nicholson, A. E., & Korb, K. B. (2015). Fog Forecasting for Melbourne Airport Using a Bayesian Decision Network. *Weather and Forecasting*, *30*(5), 1218-1233. doi: 10.1175/WAF-D-15-0005.1

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

Buntine, W. (1991). Theory Refinement on Bayesian Networks. In B. D. D'Ambrosio, P. Smets, & P. P. Bonissone (Eds.), *Uncertainty proceedings 1991* (p. 52 - 60). San Francisco (CA): Morgan Kaufmann. doi: https://doi.org/10.1016/B978-1-55860-203-8.50010-3

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

Catenacci, M., & Giupponi, C. (2009). *Potentials of Bayesian Networks to Deal with Uncertainty in Climate Change Adaptation Policies* (SSRN Scholarly Paper No. ID 1526605). Rochester, NY: Social Science Research Network. doi: 10.2139/ssrn.1526605

Catenacci, M., & Giupponi, C. (2013). Integrated assessment of sea-level rise adaptation strategies using a Bayesian decision network approach. *Environmental Modelling and Software*, *44*, 87 - 100. doi: 10.1016/j.envsoft.2012.10.010

Colombo, D., & Maathuis, M. H. (2014). Order-Independent Constraint-Based Causal Structure Learning. *Journal of Machine Learning Research*, *15*(116), 3921-3962.

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

Dawid, A. P., & Lauritzen, S. L. (1993). Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. *The Annals of Statistics*, *21*(3), 1272–1317.

Dechter, R. (1999). Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence*, *113*(1), 41 - 85. doi: https://doi.org/10.1016/S0004-3702(99)00059-4

Deng, Y., & Ebert-Uphoff, I. (2014). Weakening of atmospheric information flow in a warming climate in the community climate system model. *Geophysical Research Letters*, *41*(1), 193-200. doi: 10.1002/2013GL058646

Deser, C. (2000). On the teleconnectivity of the "Arctic Oscillation". *Geophysical Research Letters*, *27*(6), 779-782. doi: https://doi.org/10.1029/1999GL010945

Di Capua, G., Runge, J., Donner, R. V., van den Hurk, B., Turner, A. G., Vellore, R., … Coumou, D. (2020). Dominant patterns of interaction between the tropics and mid-latitudes in boreal summer: causal relationships and the role of timescales. *Weather and Climate Dynamics*, *1*(2), 519–539. doi: 10.5194/wcd-1-519-2020

Donges, J. F., Schultz, H. C. H., Marwan, N., Zou, Y., & Kurths, J. (2011). Investigating the topology of interacting networks. *The European Physical Journal B*, *84*(4), 635–651. doi: 10.1140/epjb/e2011-10795-8

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

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

Ebert-Uphoff, I., & Deng, Y. (2012b). A new type of climate network based on probabilistic graphical models: Results of boreal winter versus summer. *Geophysical Research Letters*, *39*(19). doi: 10.1029/2012GL053269

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

Eichler, M. (2012, Jun 01). Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, *153*(1), 233–268. doi: 10.1007/s00440-011-0345-8

Eyring, V., Cox, P. M., Flato, G. M., Gleckler, P. J., Abramowitz, G., Caldwell, P., … Williamson, M. S. (2019). Taking climate model evaluation to the next level. *Nature Climate Change*, *9*(2), 102–110. doi: 10.1038/s41558-018-0355-y

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

Fountalis, I., Dovrolis, C., Bracco, A., Dilkina, B., & Keilholz, S. (2018). δ-MAPS: from spatio-temporal data to a weighted and lagged network between functional domains. *Applied Network Science*, *3*(1), 21. doi: 10.1007/s41109-018-0078-z

Franzke, C., Feldstein, S. B., & Lee, S. (2011). Synoptic analysis of the Pacific–North American teleconnection pattern. *Quarterly Journal of the Royal Meteorological Society*, *137*(655), 329-346. doi: https://doi.org/10.1002/qj.768

Frederiksen, J. S., & Frederiksen, C. S. (1993). Monsoon Disturbances, Intraseasonal Oscillations, Teleconnection Patterns, Blocking, and Storm Tracks of the Global Atmosphere during January 1979: Linear Theory. *Journal of the Atmospheric Sciences*, *50*(10), 1349-1372. doi: 10.1175/1520-0469(1993)050⟨1349:MDIOTP⟩2.0.CO;2

Friedman, N., & Goldszmidt, M. (1996). Discretizing Continuous Attributes While Learning Bayesian Networks. In *ICML 1996.*

Friedman, N., Goldszmidt, M., & Wyner, A. (1999). Data Analysis with Bayesian Networks: A Bootstrap Approach. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* (p. 196–205). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

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).

Ghil, M., & Lucarini, V. (2020). The physics of climate variability and climate change. *Rev. Mod. Phys.*, *92*, 035002. doi: 10.1103/RevModPhys.92.035002

Gillett, N. P., Kell, T. D., & Jones, P. D. (2006). Regional climate impacts of the Southern Annular Mode. *Geophysical Research Letters*, *33*(23), L23704. doi: 10.1029/2006GL027721

Goddard, L., Mason, S., Zebiak, S., Ropelewski, C., Basher, R., & Cane, M. (2001). Current approaches to seasonal to interannual climate predictions. *International Journal of Climatology*, *21*(9), 1111-1152. doi: 10.1002/joc.636

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

Gozolchiani, A., Havlin, S., & Yamasaki, K. (2011). Emergence of El Niño as an Autonomous Component in the Climate Network. *Phys. Rev. Lett.*, *107*, 148501. doi: 10.1103/PhysRevLett.107.148501

Gozolchiani, A., Yamasaki, K., Gazit, O., & Havlin, S. (2008). Pattern of climate network blinking links follows El Niño events. *EPL (Europhysics Letters)*, *83*(2), 28005. doi: 10.1209/0295-5075/83/28005

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

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

Greenland, S., & Brumback, B. (2002). An overview of relations among causal modelling methods. *International Journal of Epidemiology*, *31*(5), 1030-1037. doi: 10.1093/ije/31.5.1030

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

Guez, O., Gozolchiani, A., Berezin, Y., Brenner, S., & Havlin, S. (2012). Climate network structure evolves with North Atlantic Oscillation phases. *EPL (Europhysics Letters)*, *98*(3), 38006. doi: 10.1209/0295-5075/98/38006

Hannachi, A., Straus, D. M., Franzke, C. L. E., Corti, S., & Woollings, T. (2017). Low-frequency nonlinearity and regime behavior in the Northern Hemisphere extratropical atmosphere. *Reviews of Geophysics*, *55*(1), 199-234. doi: 10.1002/2015RG000509

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

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 Presatellite 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

Hlinka, J., Hartman, D., Vejmelka, M., Runge, J., Marwan, N., Kurths, J., & Paluš, M. (2013). Reliability of Inference of Directed Climate Networks Using Conditional Mutual Information. *Entropy*, *15*(6), 2023–2045. doi: 10.3390/e15062023

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

Horenko, I., Gerber, S., O'Kane, T. J., Risbey, J. S., & Monselesan, D. P. (2017). On inference and validation of causality relations in climate teleconnections. In C. L. E. Franzke & T. J. O'Kane (Eds.), *Nonlinear and stochastic climate dynamics* (pp. 184–208). Cambridge University Press.

Hoskins, B. J., & Ambrizzi, T. (1993). Rossby Wave Propagation on a Realistic Longitudinally Varying Flow. *Journal of the Atmospheric Sciences*, *50*(12), 1661-1671. doi: 10.1175/1520-0469(1993)050⟨1661:RWPOAR⟩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

Hurrell, J. W., Kushnir, Y., Ottersen, G., & Visbeck, M. (Eds.). (2003). *The North Atlantic Oscillation: Climatic Significance and Environmental Impact* (Vol. 134). American Geophysical Union.

Jordan, M. I. (2004). Graphical models. *Statist. Sci.*, *19*(1), 140–155. doi: 10.1214/088342304000000026

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

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

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models.* Cambridge: The MIT Press.

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

10.1175/JCLI-D-15-0654.1

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., & Phillips, T. J. (1986). Coherent Fluctuations of Extratropical Geopotential Height and Tropical Convection in Intraseasonal Time Scales. *Journal of the Atmospheric Sciences*, *43*(11), 1164-1181. doi: 10.1175/1520-0469(1986)043⟨1164:CFOFGH⟩2.0.CO;2

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

Leathers, D. J., Yarnal, B., & Palecki, M. A. (1991). The Pacific/North American Teleconnection Pattern and United States Climate. Part I: Regional Temperature and Precipitation Associations. *Journal of Climate*, *4*(5), 517-528. doi: 10.1175/1520-0442(1991)004⟨0517:TPATPA⟩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

Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., & Schnellnhuber, H. J. (2008). Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences*, *105*(6), 1786–1793. doi: 10.1073/pnas.0705414105

Leonard, M., Westra, S., Phatak, A., Lambert, M., van den Hurk, B., McInnes, K., … Stafford-Smith, M. (2014). A compound event framework for understanding extreme impacts. *WIREs Climate Change*, *5*(1), 113-128. doi: 10.1002/wcc.252

Li, S., Wang, M., Bond, N. A., Huang, W., Wang, Y., Xu, S., … Bai, Y. (2018). Precursors of September Arctic Sea-Ice Extent Based on Causal Effect Networks. *Atmosphere*, *9*(11), 437. doi: 10.3390/atmos9110437

Lin, H., Brunet, G., & Derome, J. (2009). An Observed Connection between the

North Atlantic Oscillation and the Madden–Julian Oscillation. *Journal of Climate*, *22*(2), 364-380. doi: 10.1175/2008JCLI2515.1

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.

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

Maloney, E. D., Gettelman, A., Ming, Y., Neelin, J. D., Barrie, D., Mariotti, A., ... Zhao, M. (2019). Process-Oriented Evaluation of Climate and Weather Forecasting Models. *Bulletin of the American Meteorological Society*, *100*(9), 1665-1686. doi: 10.1175/BAMS-D-18-0042.1

Marshall, A. G., Hudson, D., Wheeler, M. C., Alves, O., Hendon, H. H., Pook, M. J., & Risbey, J. S. (2014). Intra-seasonal drivers of extreme heat over Australia in observations and POAMA-2. *Climate Dynamics*, *43*(7), 1915–1937. doi: 10.1007/s00382-013-2016-1

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

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

McPhaden, M. J., Santoso, A., & Cai, W. (Eds.). (2021). *El Niño Southern Oscilla-*

tion in a Changing Climate (Vol. 253). American Geophysical Union.

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

Mo, K. C., & Paegle, J. N. (2001). The Pacific–South American modes and their downstream effects. *International Journal of Climatology*, *21*(10), 1211-1229. doi: 10.1002/joc.685

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., Horenko, I., & Franzke, C. L. (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.

Pearl, J. (1982). Reverend Bayes on Inference Engines: A Distributed Hierarchical Approach. In *Proceedings of the Second National Conference on Artificial Intelligence (AAAI 1982)* (p. 133-136). AAAI Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems.* San Francisco, California: Morgan Kaufmann Publishers, Inc.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669-688. doi: 10.1093/biomet/82.4.669

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.

Peter, C., Lange, W. d., Musango, J. K., April, K., & Potgieter, A. (2009). Applying Bayesian modelling to assess climate change effects on biofuel production. *Climate Research*, *40*(2-3), 249–260. doi: 10.3354/cr00833

Pfleiderer, P., Schleussner, C.-F., Geiger, T., & Kretschmer, M. (2020). Robust predictors for seasonal Atlantic hurricane activity identified with causal effect networks. *Weather and Climate Dynamics*, *1*(2), 313–324. doi: 10.5194/wcd-1-313-2020

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

Radebach, A., Donner, R. V., Runge, J., Donges, J. F., & Kurths, J. (2013). Disentangling different types of El Niño episodes by evolving climate network analysis. *Phys. Rev. E*, *88*, 052807. doi: 10.1103/PhysRevE.88.052807

Ångström, A. (1935). Teleconnections of climatic changes in present time. *Geografiska Annaler*, *17*(3-4), 242-258. doi: 10.1080/20014422.1935.11880600

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. (2018a). 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. (2018b). Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In A. Storkey & F. Perez-Cruz (Eds.), *Proceedings of the Twenty-First International Conference on Artificial*

Intelligence and Statistics (Vol. 84, pp. 938–947). Playa Blanca, Lanzarote, Canary Islands: PMLR.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., ... Zscheischler, J. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, *10*, 2553. doi: 10.1038/s41467-019-10105-3

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

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

Saggioro, E., de Wiljes, J., Kretschmer, M., & Runge, J. (2020). *Reconstructing regime-dependent causal relationships from observational time series.* (arXiv:2007.00267)

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

Samarasinghe, S. M., Deng, Y., & Ebert-Uphoff, I. (2020). A Causality-Based View of the Interaction between Synoptic- and Planetary-Scale Atmospheric Disturbances. *Journal of the Atmospheric Sciences*, *77*(3), 925-941. doi: 10.1175/JAS-D-18-0163.1

Samarasinghe, S. M., McGraw, M. C., Barnes, E. A., & Ebert-Uphoff, I. (2019). A study of links between the Arctic and the midlatitude jet stream using Granger and Pearl causality. *Environmetrics*, *30*(4), e2540. doi: 10.1002/env.2540

Schott, F. A., Xie, S.-P., & McCreary Jr., J. P. (2009). Indian ocean circulation and climate variability. *Reviews of Geophysics*, *47*(1), RG1002. doi: https://doi.org/10.1029/2007RG000245

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

Sisson, S. A. (2005). Transdimensional Markov Chains. *Journal of the American Statistical Association*, *100*(471), 1077-1089. doi: 10.1198/016214505000000664

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., & Cowell, R. G. (1993). Bayesian Analysis in Expert Systems. *Statistical Science*, *8*(3), 219–247.

Spiegelhalter, D. J., & Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, *20*(5), 579-605. doi: 10.1002/net.3230200507

Spirtes, P., & Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, *9*(1), 62–72.

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

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.

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

Trenberth, K. E., Branstator, G. W., Karoly, D., Kumar, A., Lau, N.-C., & Ropelewski, C. (1998). Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *Journal of Geophysical Research: Oceans*, *103*(C7), 14291-14324. doi:

10.1029/97JC01444

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., & Kravtsov, S. (2007). A new dynamical mechanism for major climate shifts. *Geophysical Research Letters*, *34*(13). doi: 10.1029/2007GL030288

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., & Roebber, P. J. (2006). What Do Networks Have to Do with Climate? *Bulletin of the American Meteorological Society*, *87*(5), 585–596. (Publisher: American Meteorological Society) doi: 10.1175/BAMS-87-5-585

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

Uusitalo, L. (2007). Advantages and challenges of Bayesian networks in environmental modelling. *Ecological Modelling*, *203*(3), 312 - 318. doi: 10.1016/j.ecolmodel.2006.11.033

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

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

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

Wang, G., Swanson, K. L., & Tsonis, A. A. (2009). The pacemaker of major climate shifts. *Geophysical Research Letters*, *36*(7). doi: 10.1029/2008GL036874

Wang, Y., Gozolchiani, A., Ashkenazy, Y., Berezin, Y., Guez, O., & Havlin, S. (2013). Dominant Imprint of Rossby Waves in the Climate Network. *Phys. Rev. Lett.*, *111*, 138501. doi: 10.1103/PhysRevLett.111.138501

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

Yamasaki, K., Gozolchiani, A., & Havlin, S. (2008). Climate Networks around the Globe are Significantly Affected by El Niño. *Phys. Rev. Lett.*, *100*, 228501. doi: 10.1103/PhysRevLett.100.228501

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

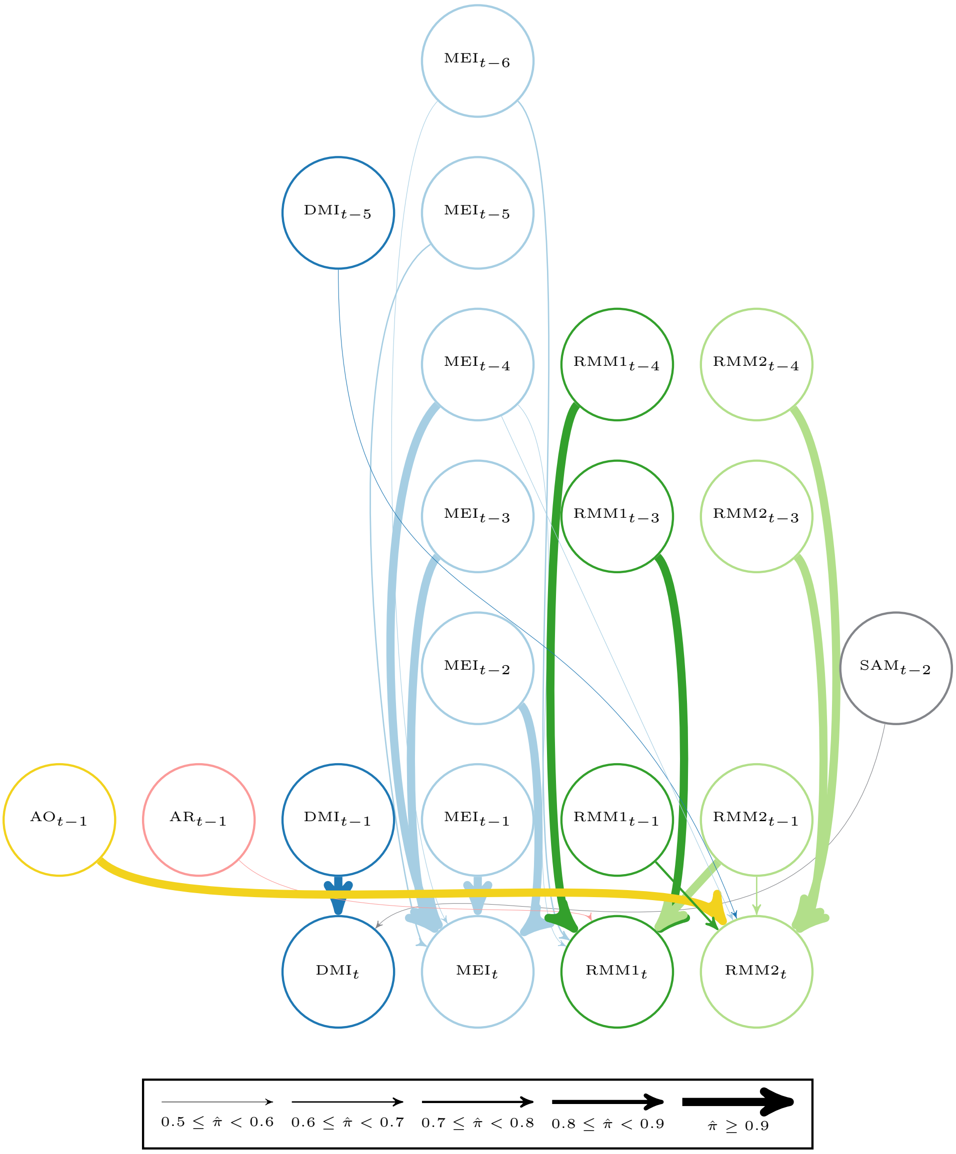
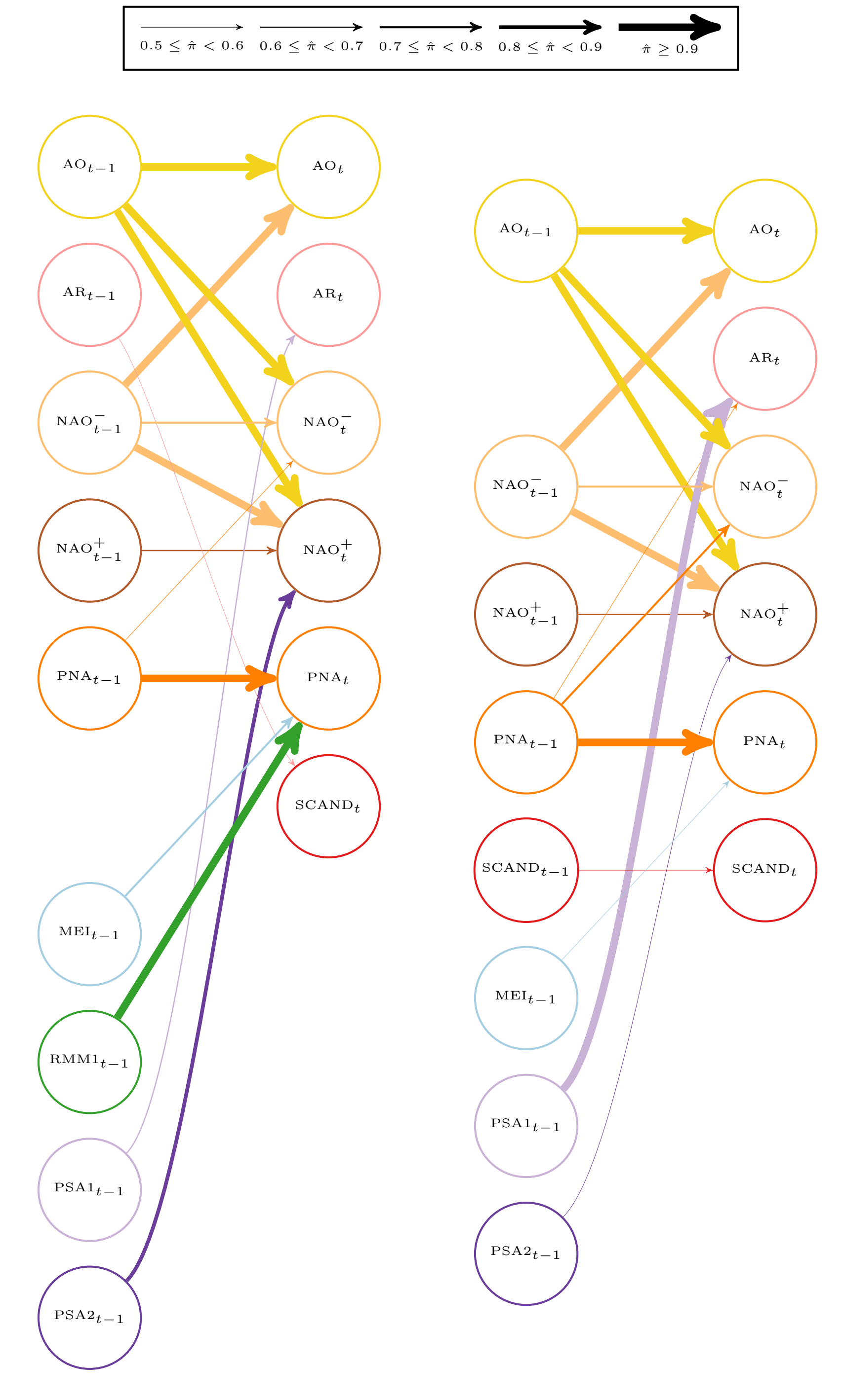


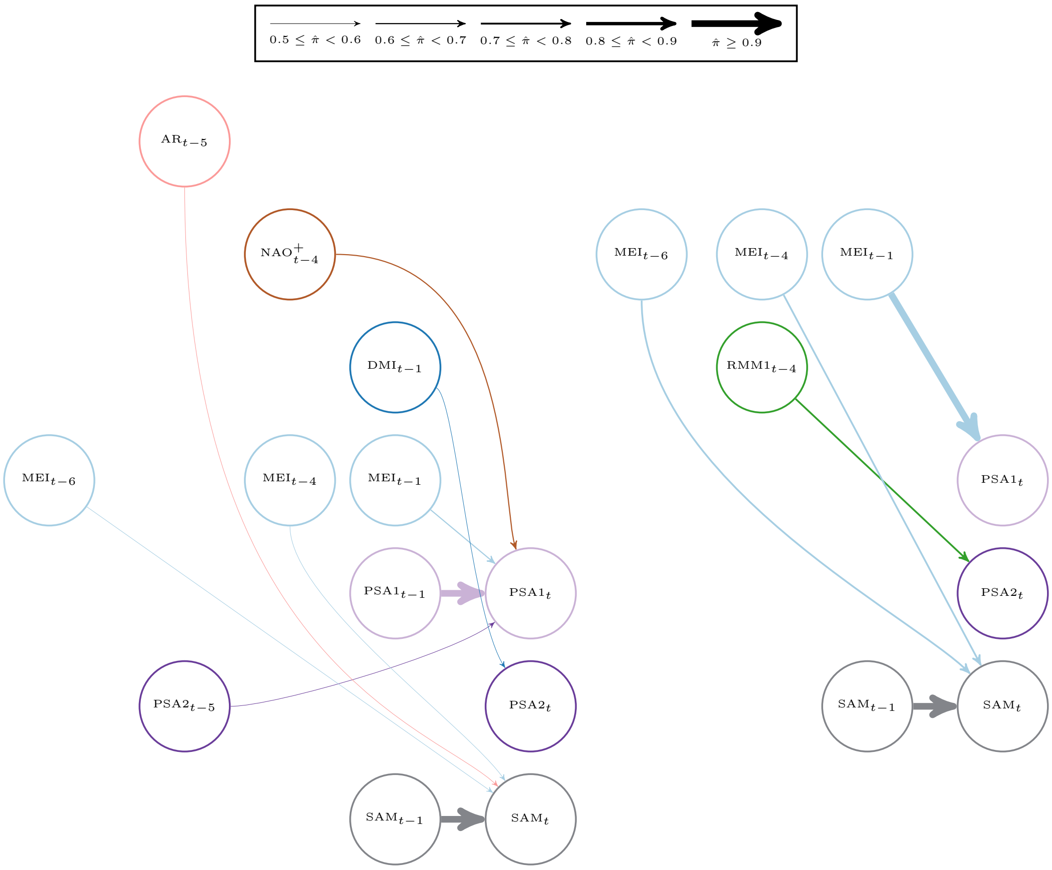
Figure 3. Subgraphs corresponding to the fitted parent sets of the tropical indices in JRA-55 based on full-year data for $a_\tau = 1.5$, $b_\tau = 20$, and $\nu^2 = 3$. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.



(a) NNR1 NH extratropical, all seasons

(b) JRA-55 NH extratropical, all seasons

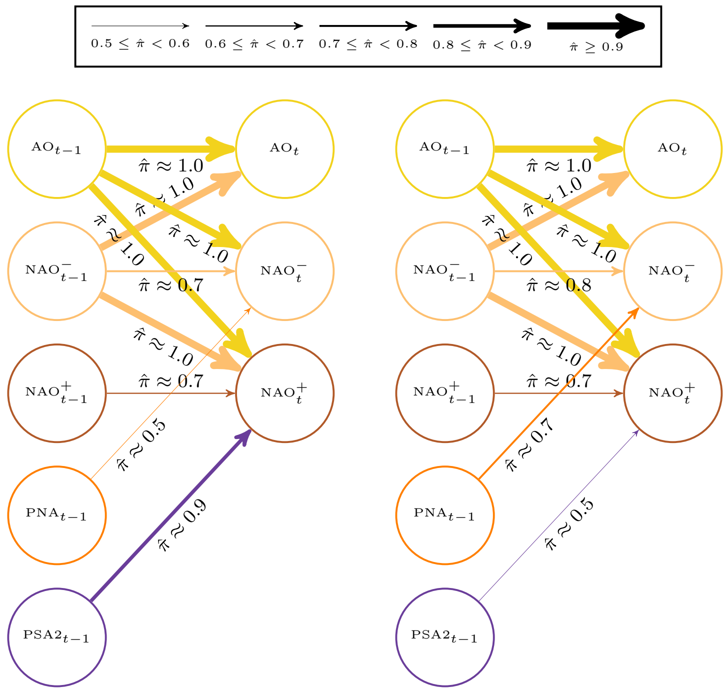
Figure 4. Subgraphs corresponding to the fitted parent sets of the NH extratropical indices in (a) NNR1 and (b) JRA-55 for $a_\tau = 1.5$, $b_\tau = 20$, and $\nu^2 = 3$. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.



(a) NNR1 SH extratropical, all seasons

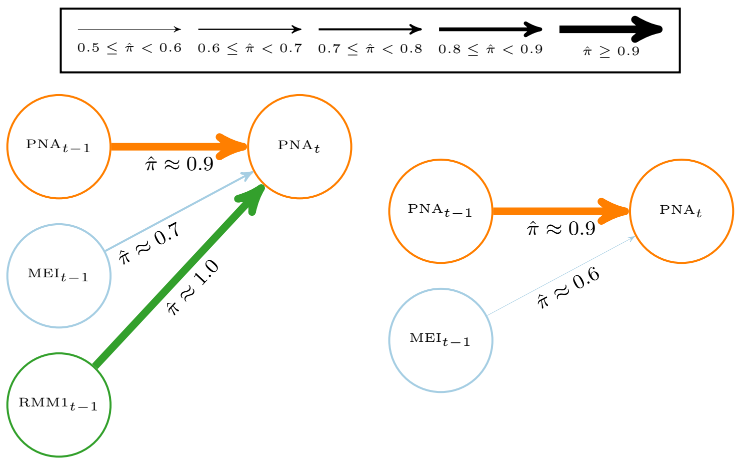
(b) JRA-55 SH extratropical, all seasons

Figure 5. Subgraphs corresponding to the fitted parent sets of the SH extratropical indices in (a) NNR1 and (b) JRA-55 for $a_\tau = 1.5$, $b_\tau = 20$, and $\nu^2 = 3$. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.



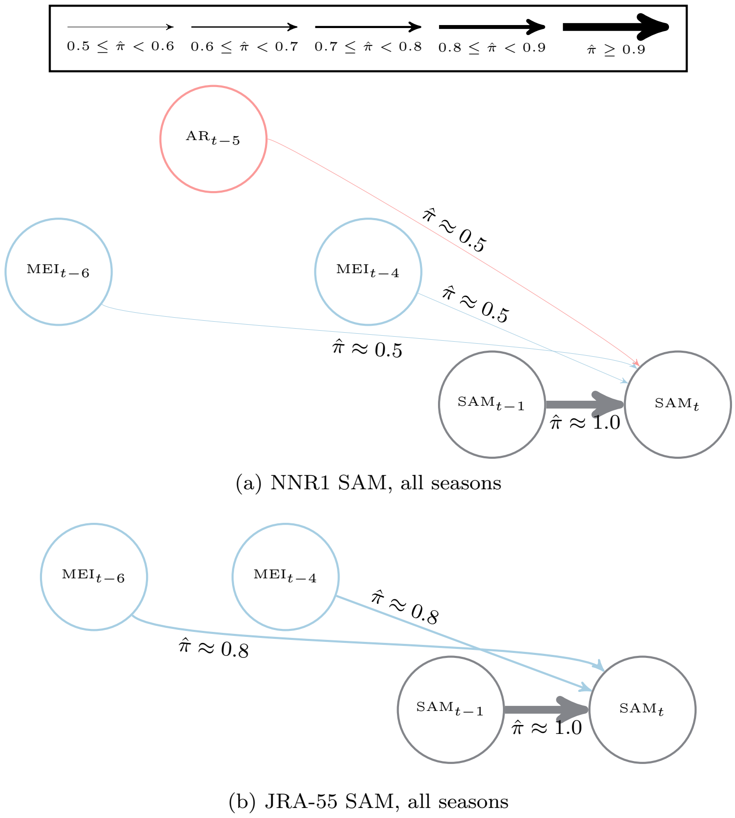
(a) NNR1 AO and NAO, all seasons (b) JRA-55 AO and NAO, all seasons

Figure 6. Subgraphs corresponding to the fitted parent set of the monthly AO and NAO indices in (a) NNR1 and (b) JRA-55, for $a_\tau = 1.5$, $b_\tau = 20$, and $\nu^2 = 3$. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.



(a) NNR1 PNA, all seasons (b) JRA-55 PNA, all seasons

Figure 7. Subgraphs corresponding to the fitted parent set of the monthly PNA index in (a) NNR1 and (b) JRA-55, for $a_\tau = 1.5$, $b_\tau = 20$, and $\nu^2 = 3$. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.



(a) NNR1 SAM, all seasons

(b) JRA-55 SAM, all seasons

Figure 8. Subgraphs corresponding to the fitted parent set of the monthly SAM index in (a) NNR1 and (b) JRA-55, for $a_\tau = 1.5$, $b_\tau = 20$, and $\nu^2 = 3$. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.

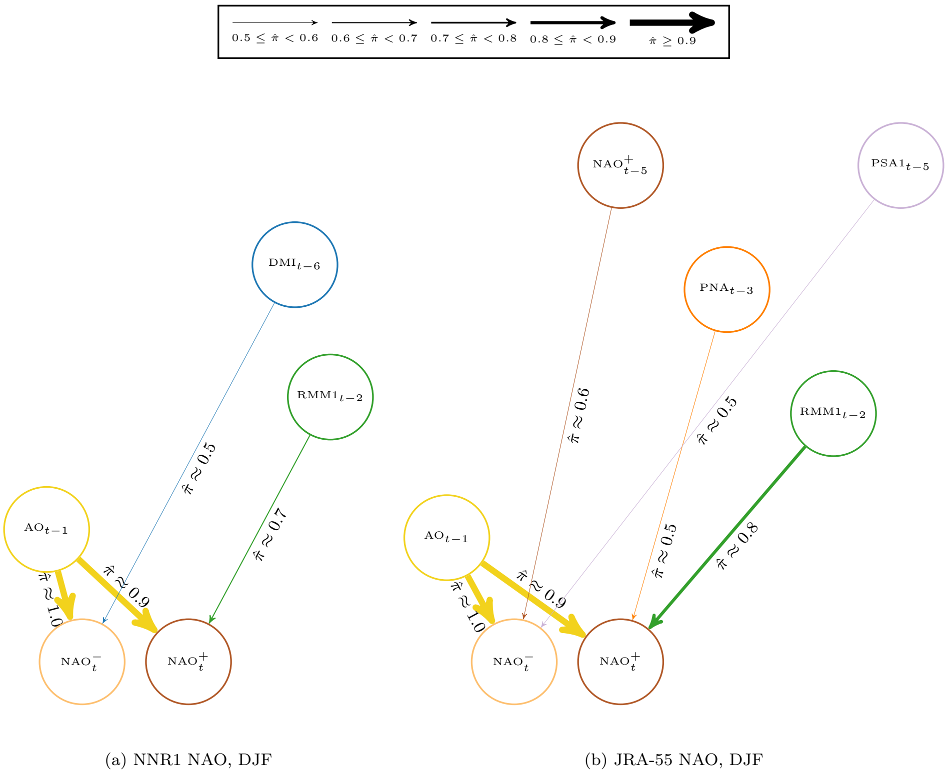
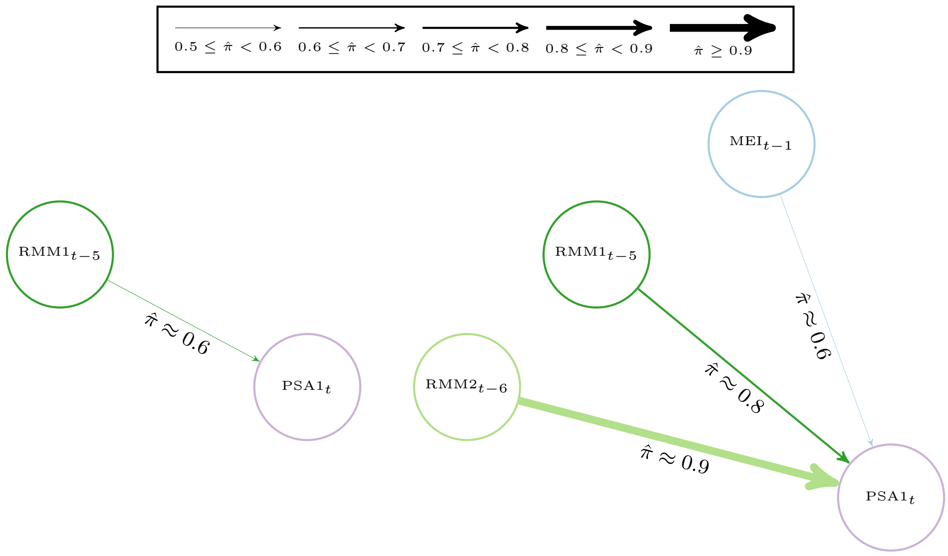


Figure 9. Subgraphs corresponding to the fitted parent set of the monthly NAO indices in (a) NNR1 and (b) JRA-55, for $a_\tau = 1.5$, $b_\tau = 20$, and $\nu^2 = 3$ during DJF. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.



(a) NNR1 PSA1, JJA (b) JRA-55 PSA1, JJA

Figure 10. Subgraphs corresponding to the fitted parent set of the monthly PSA1 index in (a) NNR1 and (b) JRA-55, for $a_\tau = 1.5$, $b_\tau = 20$, and $\nu^2 = 3$ during JJA. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.

Supporting Information for "Dynamic Bayesian networks for evaluation of Granger causal relationships in climate reanalyses"

Dylan Harries[1]and Terence J. O'Kane[1]

[1]CSIRO Oceans and Atmosphere, Hobart, Australia

Contents of this file

1. Figures S1 to S15
2. Tables S1 to S2

Introduction

This supporting information provides the patterns corresponding to the teleconnection indices analyzed in the main text, summaries of the edge posterior distributions for the alternative choice of hyperparameters, and summaries of the maximum a posteriori (MAP) structure estimates for the fits presented in the main text.

Corresponding author: Dylan Harries, CSIRO Oceans and Atmosphere, Hobart, Australia. (Dylan.Harries@csiro.au)

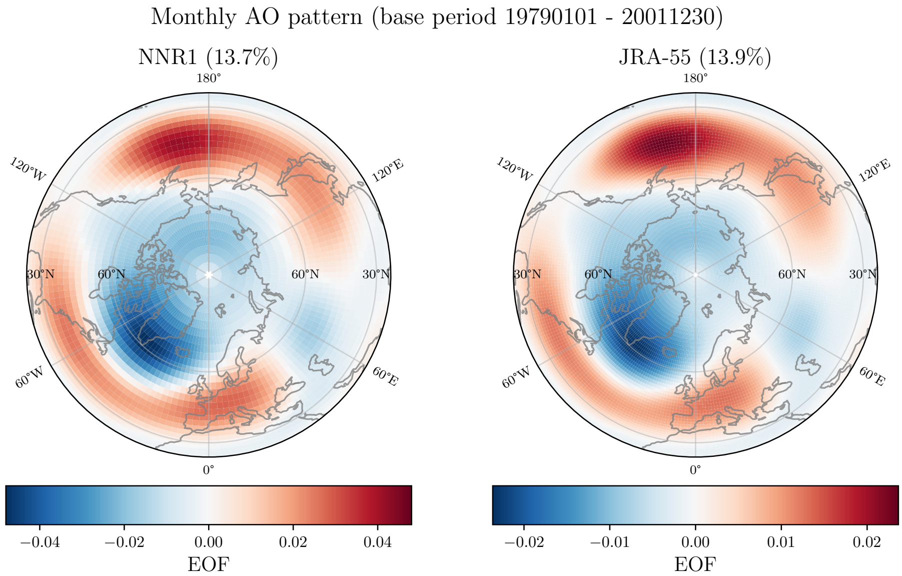


Figure S1. AO loading pattern of 500 hPa geopotential height anomalies in NNR1 (left) and JRA-55 (right).

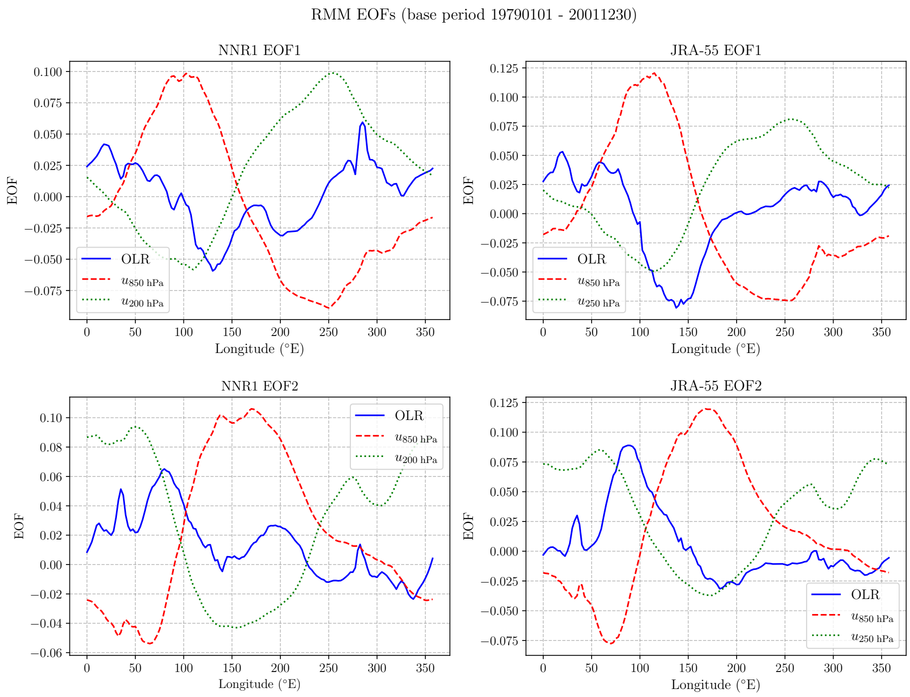


Figure S2. RMM EOFs in HadISST and NNR1 (left column) and JRA-55 (right column).

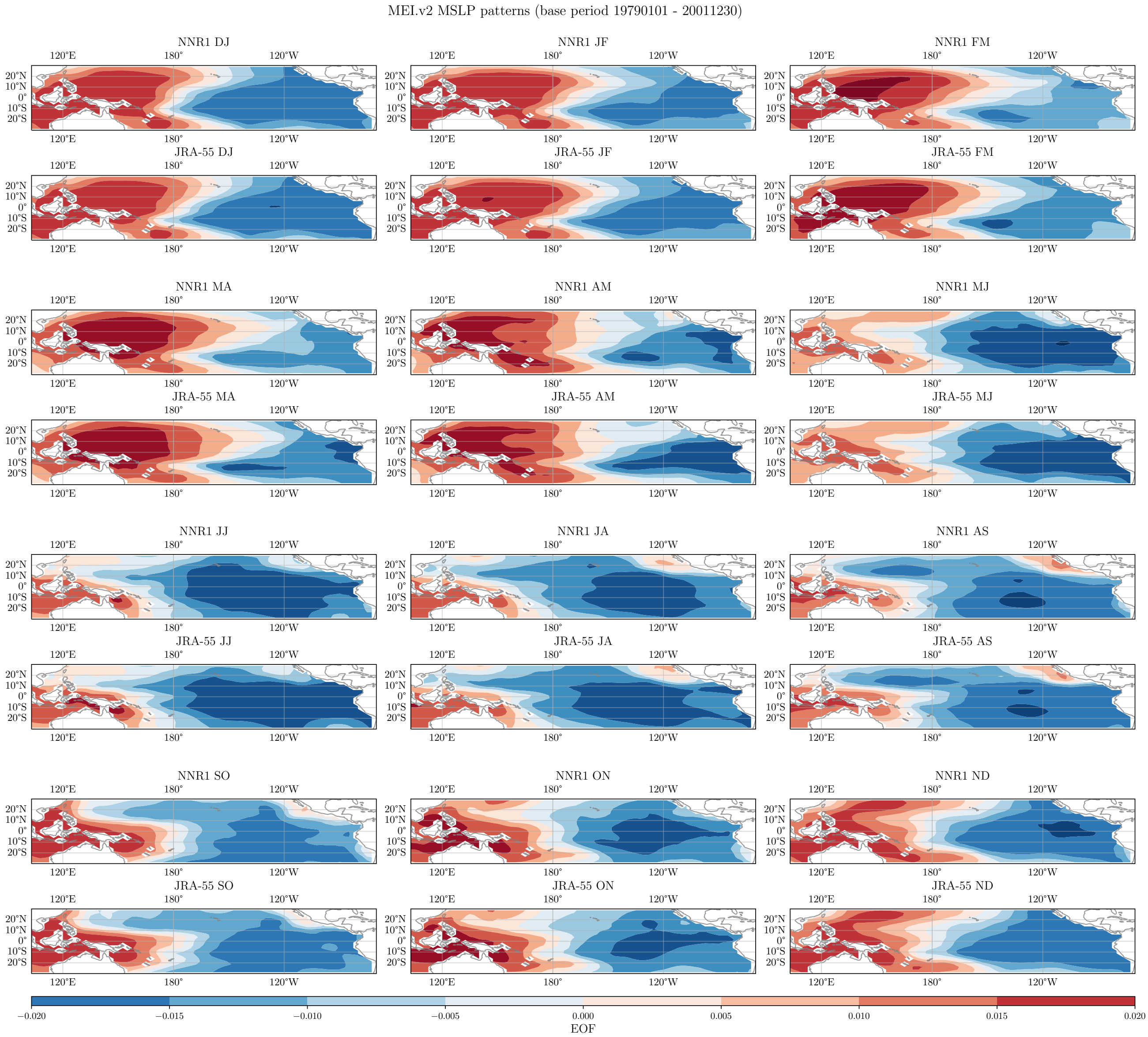


Figure S3. Seasonal MSLP anomaly patterns contributing to the MEI in HadISST and NNR1 (odd numbered rows) and JRA-55 (even numbered rows).

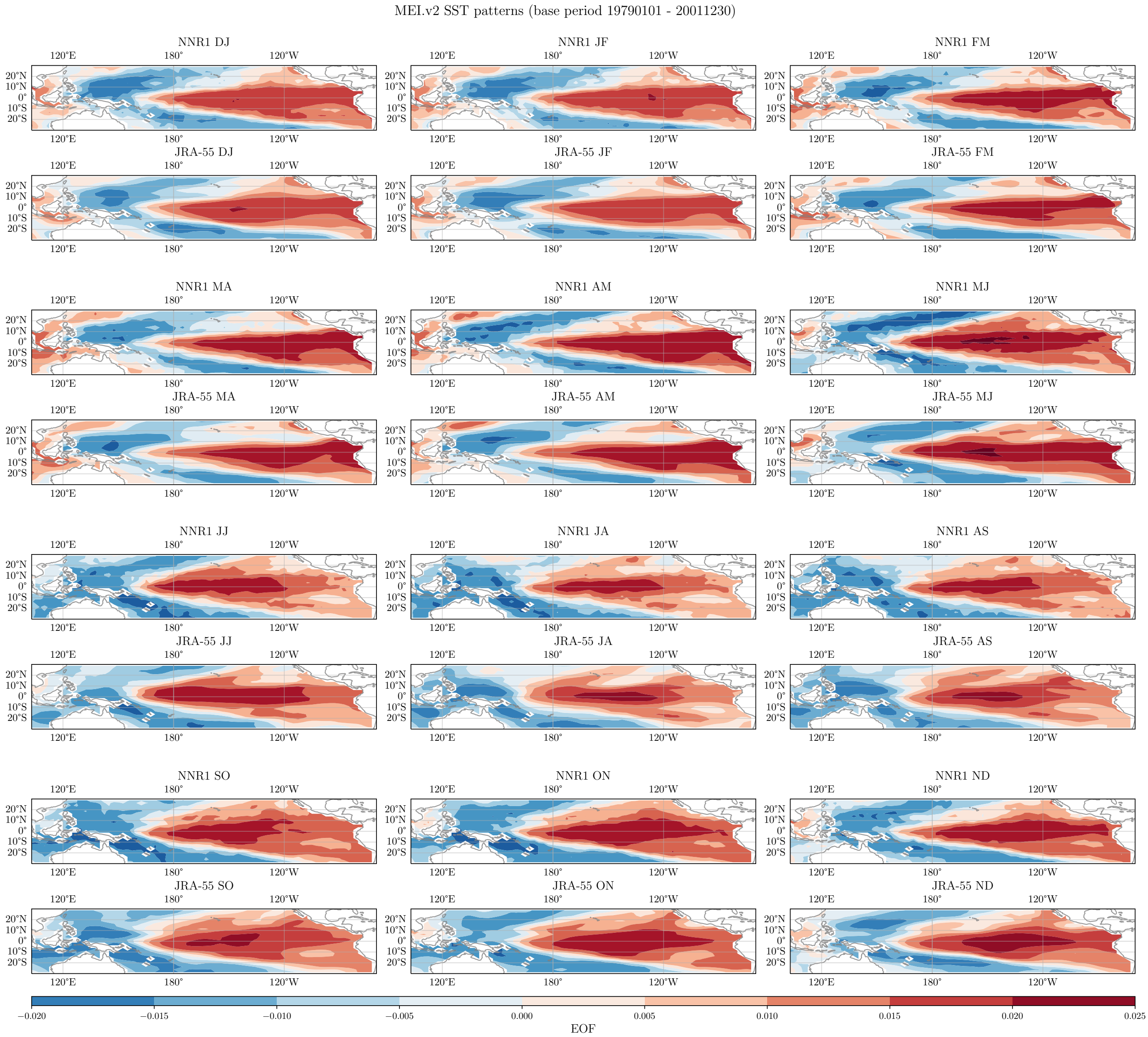


Figure S4. Seasonal SST anomaly patterns contributing to the MEI in HadISST and NNR1 (odd numbered rows) and JRA-55 (even numbered rows).

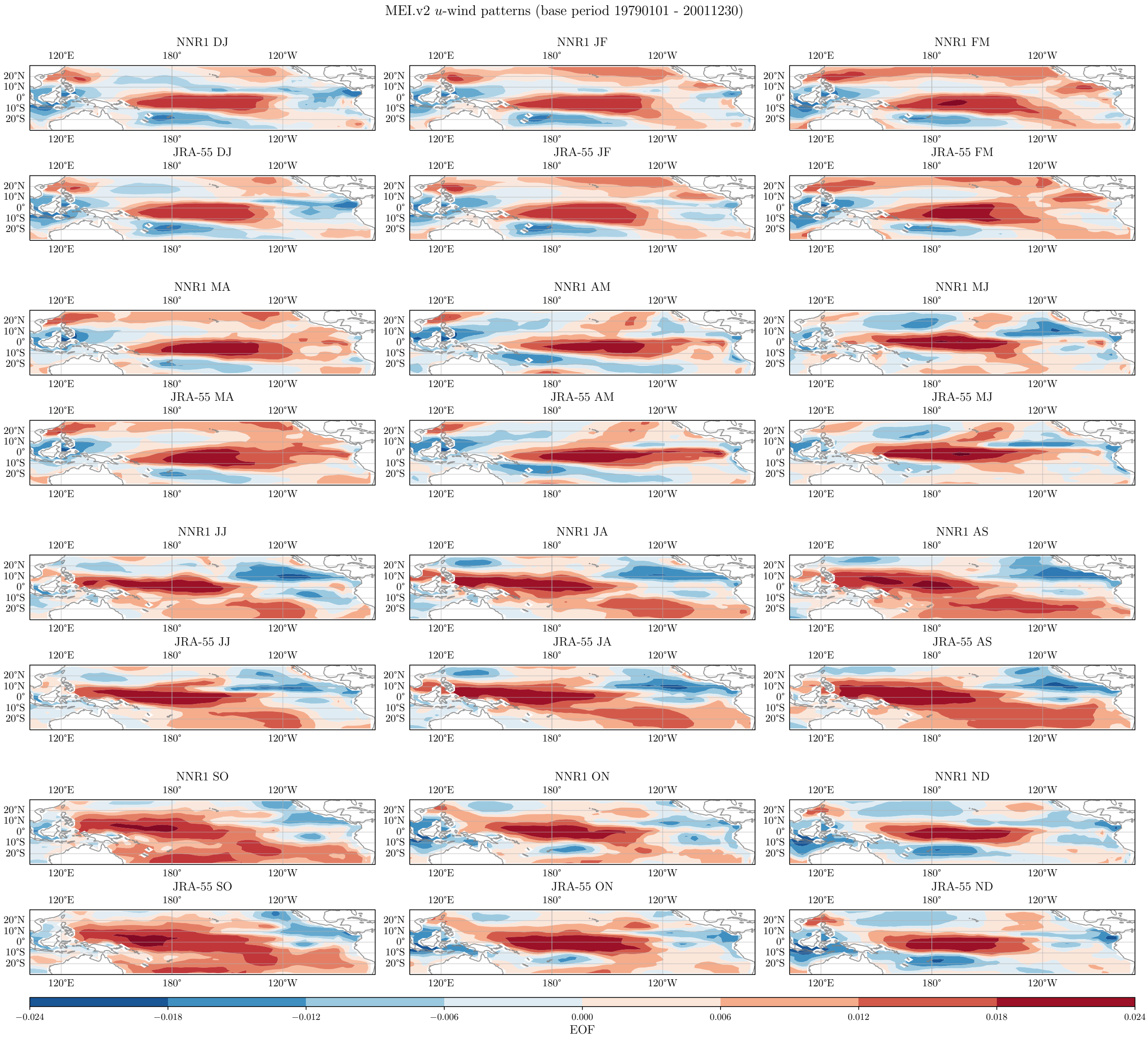


Figure S5. Seasonal zonal wind anomaly patterns contributing to the MEI in HadISST and NNR1 (odd numbered rows) and JRA-55 (even numbered rows).

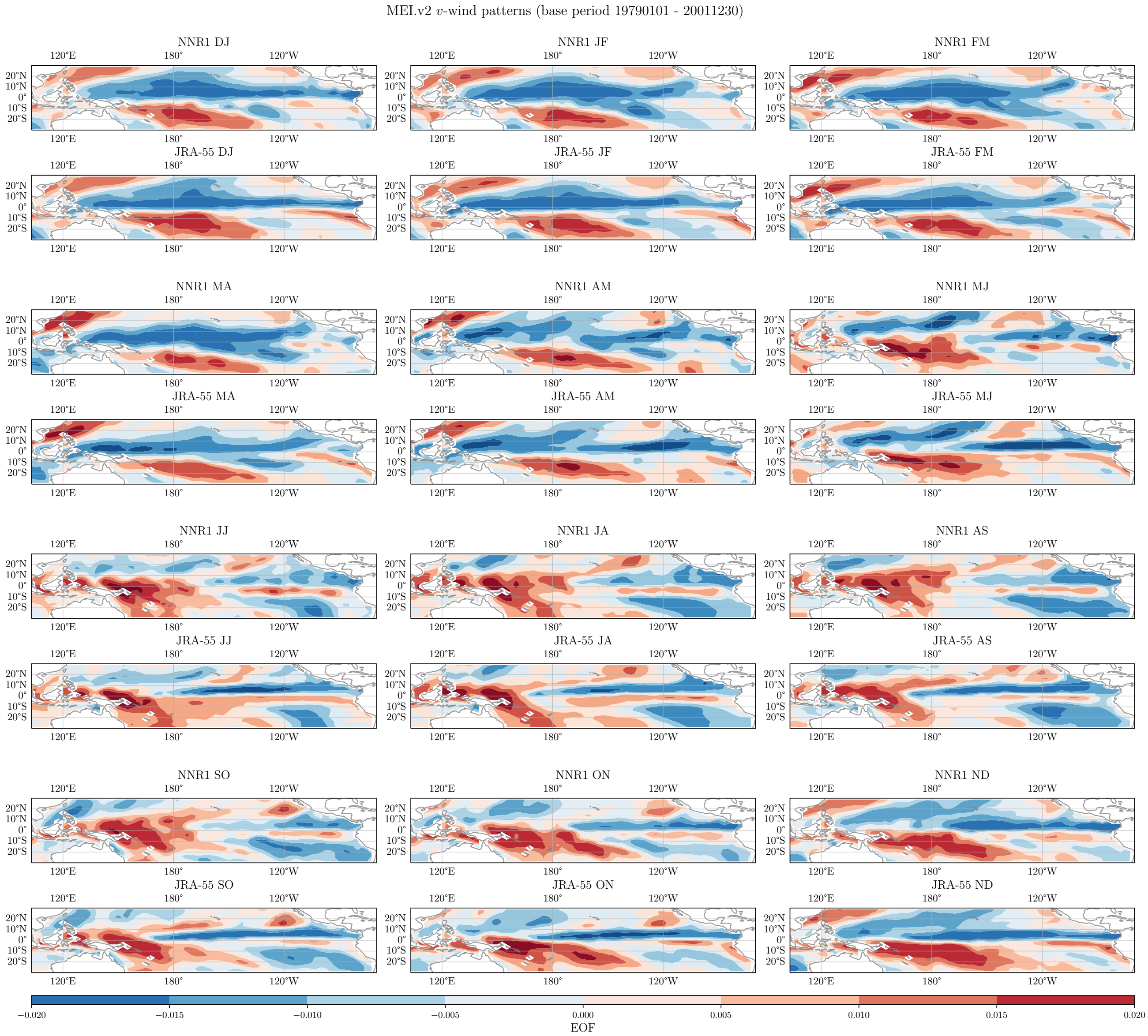


Figure S6. Seasonal meridional wind anomaly patterns contributing to the MEI in HadISST and NNR1 (odd numbered rows) and JRA-55 (even numbered rows).

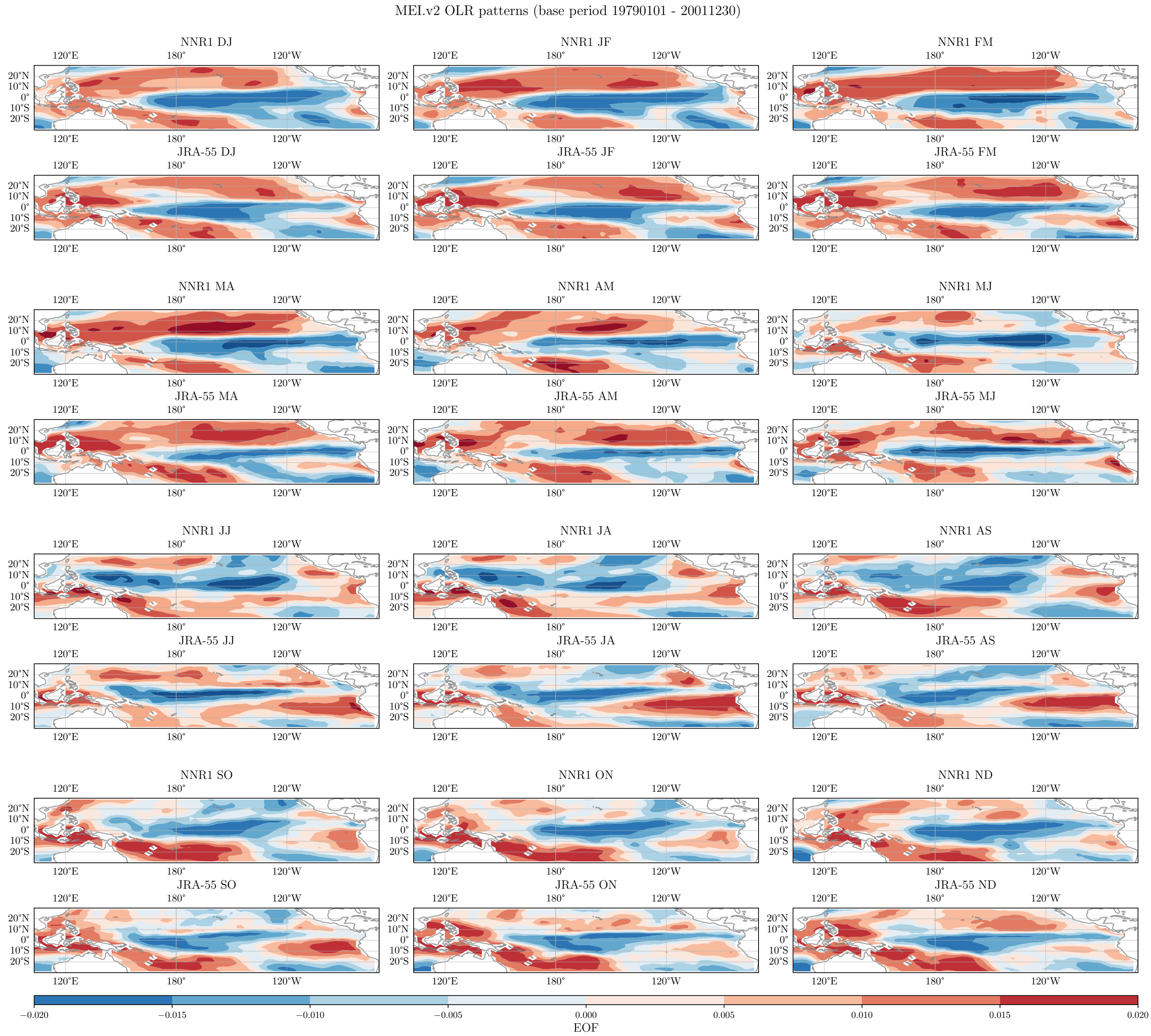


Figure S7. Seasonal OLR anomaly patterns contributing to the MEI in HadISST and NNR1 (odd numbered rows) and JRA-55 (even numbered rows).

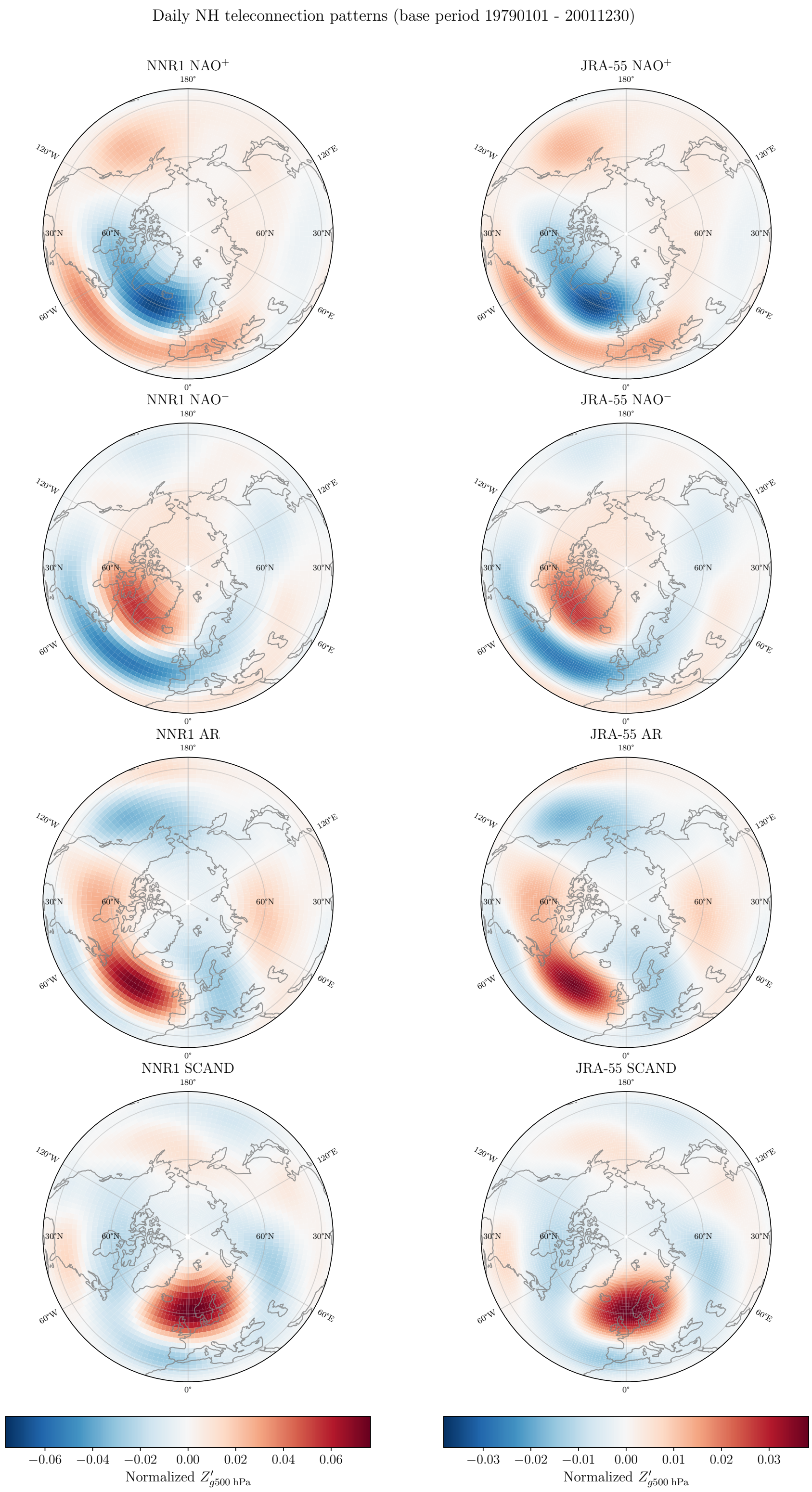


Figure S8. Northern hemisphere cluster loading patterns of 500 hPa geopotential height anomalies in NNR1 (left column) and JRA-55 (right column).

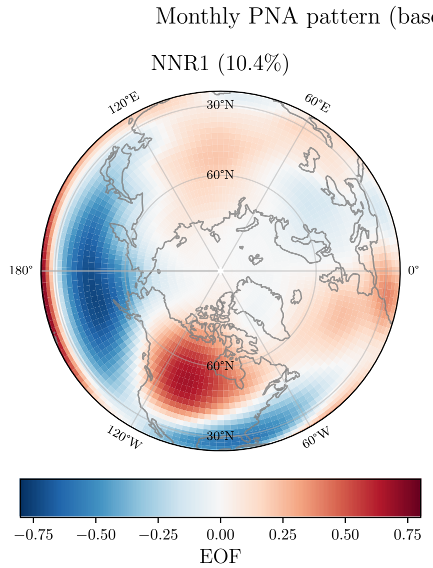


Figure S9. PNA loading pattern of 500 hPa geopotential height anomalies in NNR1 (left) and JRA-55 (right).

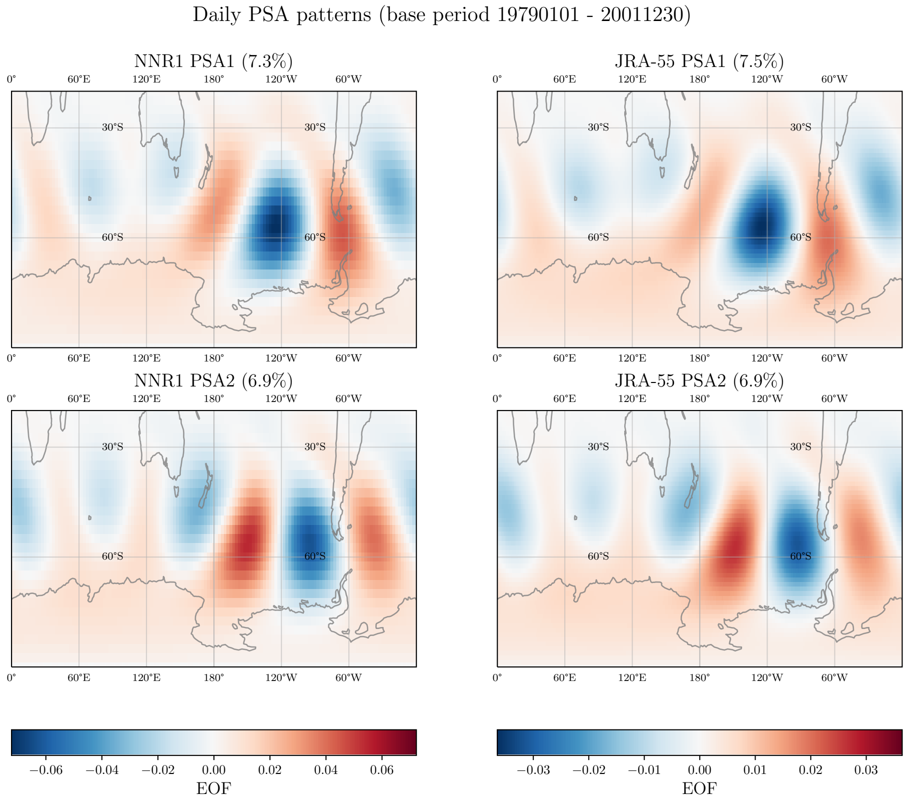


Figure S10. PSA loading patterns of 500 hPa geopotential height anomalies in NNR1 (left) and JRA-55 (right).

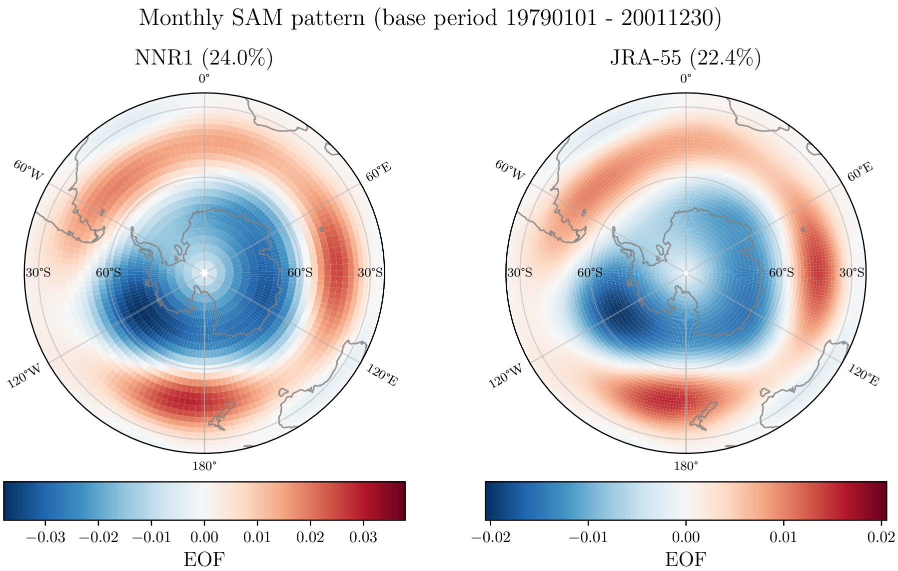


Figure S11. SAM loading pattern of 500 hPa geopotential height anomalies in NNR1 (left) and JRA-55 (right).

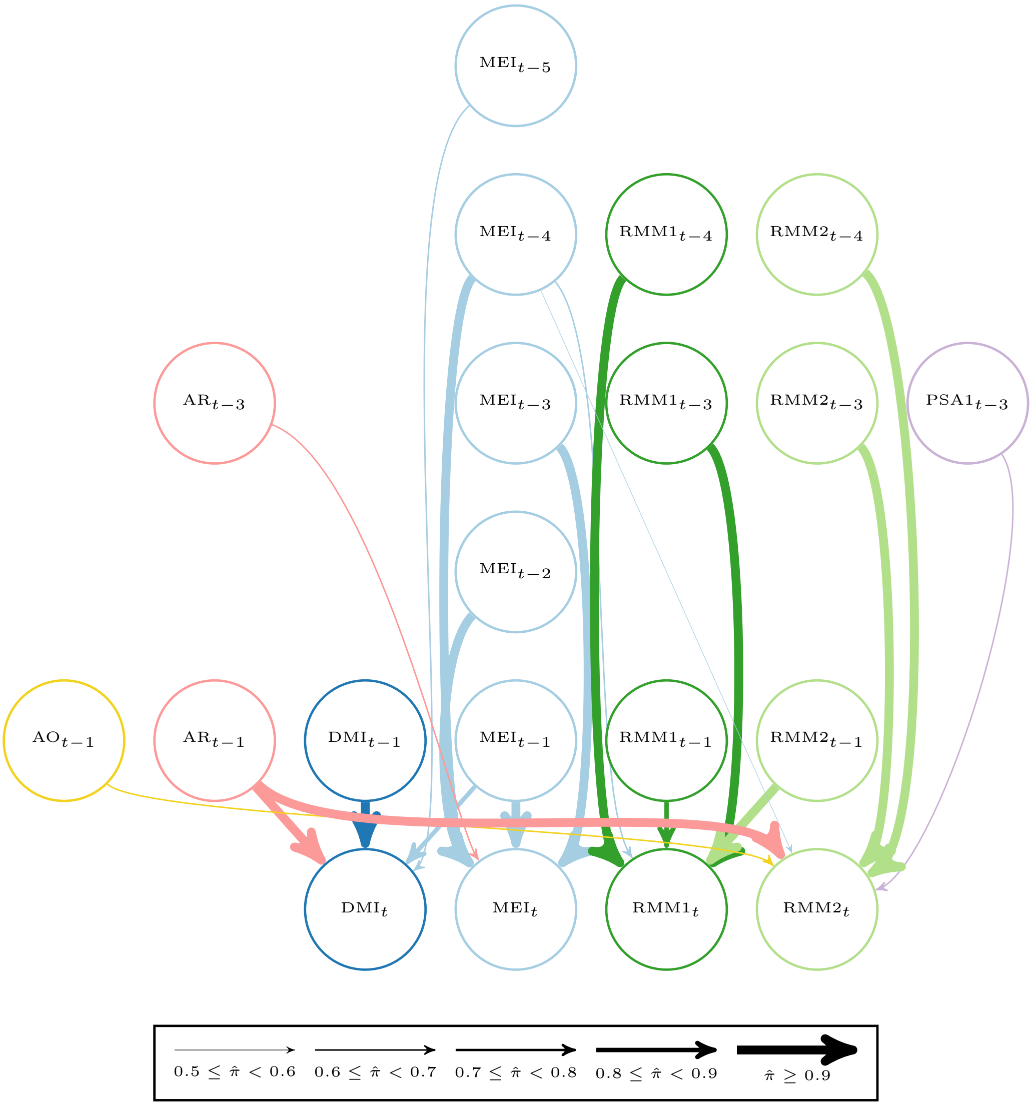


Figure S12. Subgraphs corresponding to the fitted parent sets of the tropical indices in NNR1 based on full-year data for $a_\tau = 0.5$, $b_\tau = 10$, and $\nu^2 \approx 2$. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.

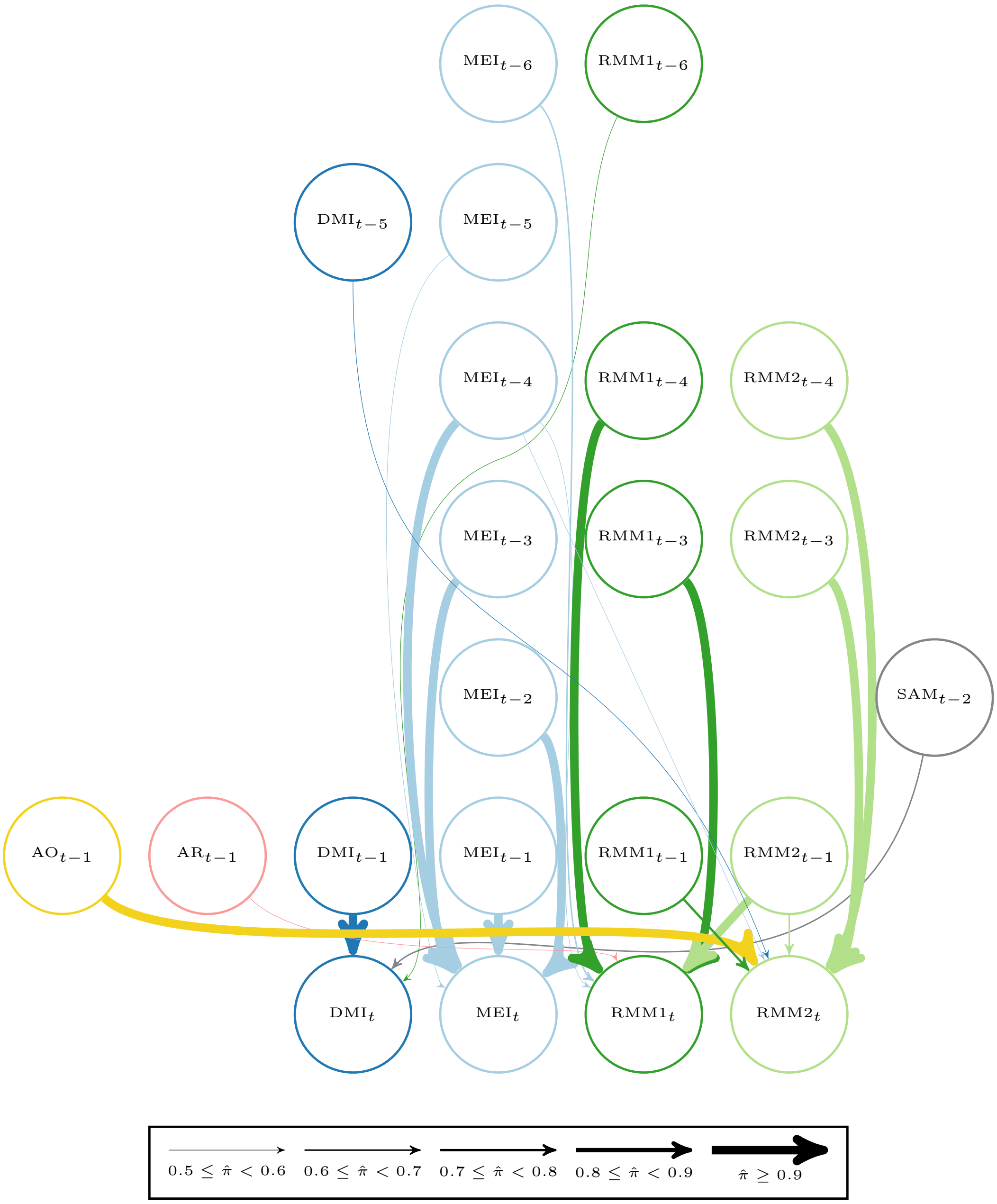
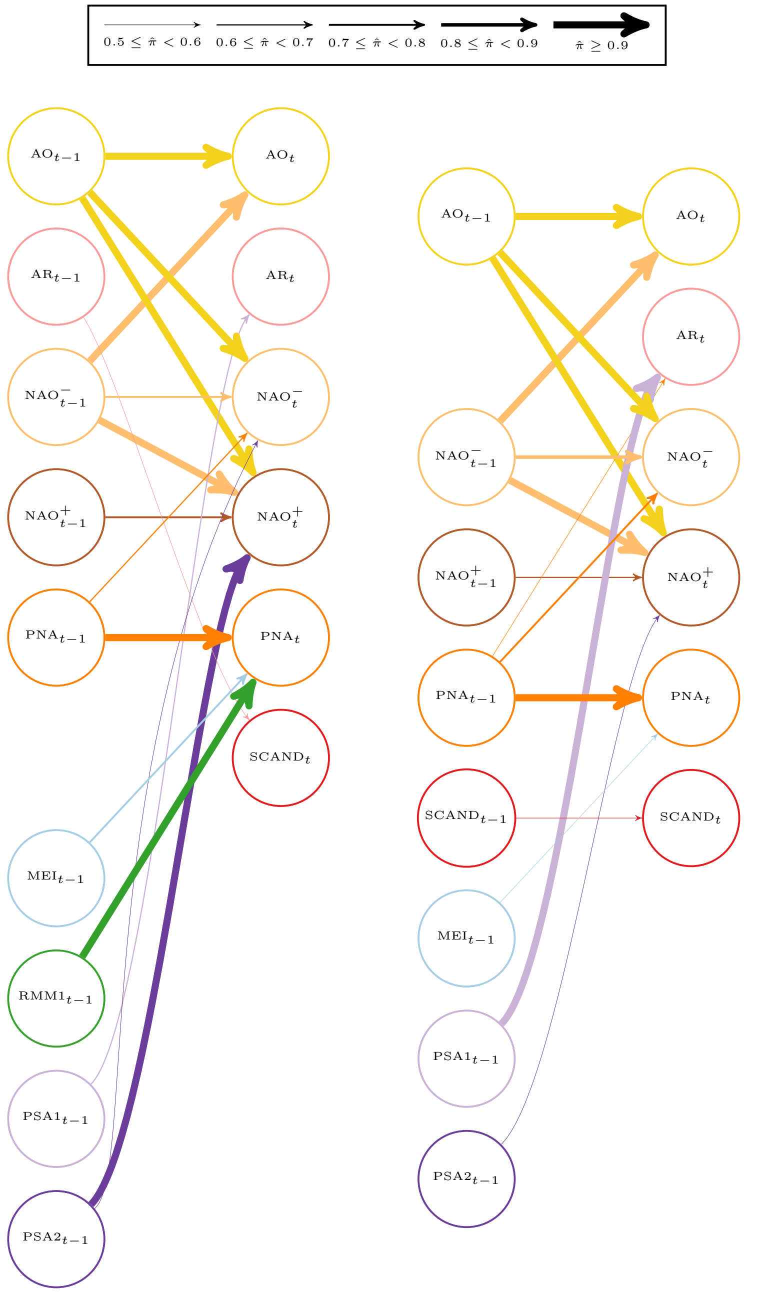
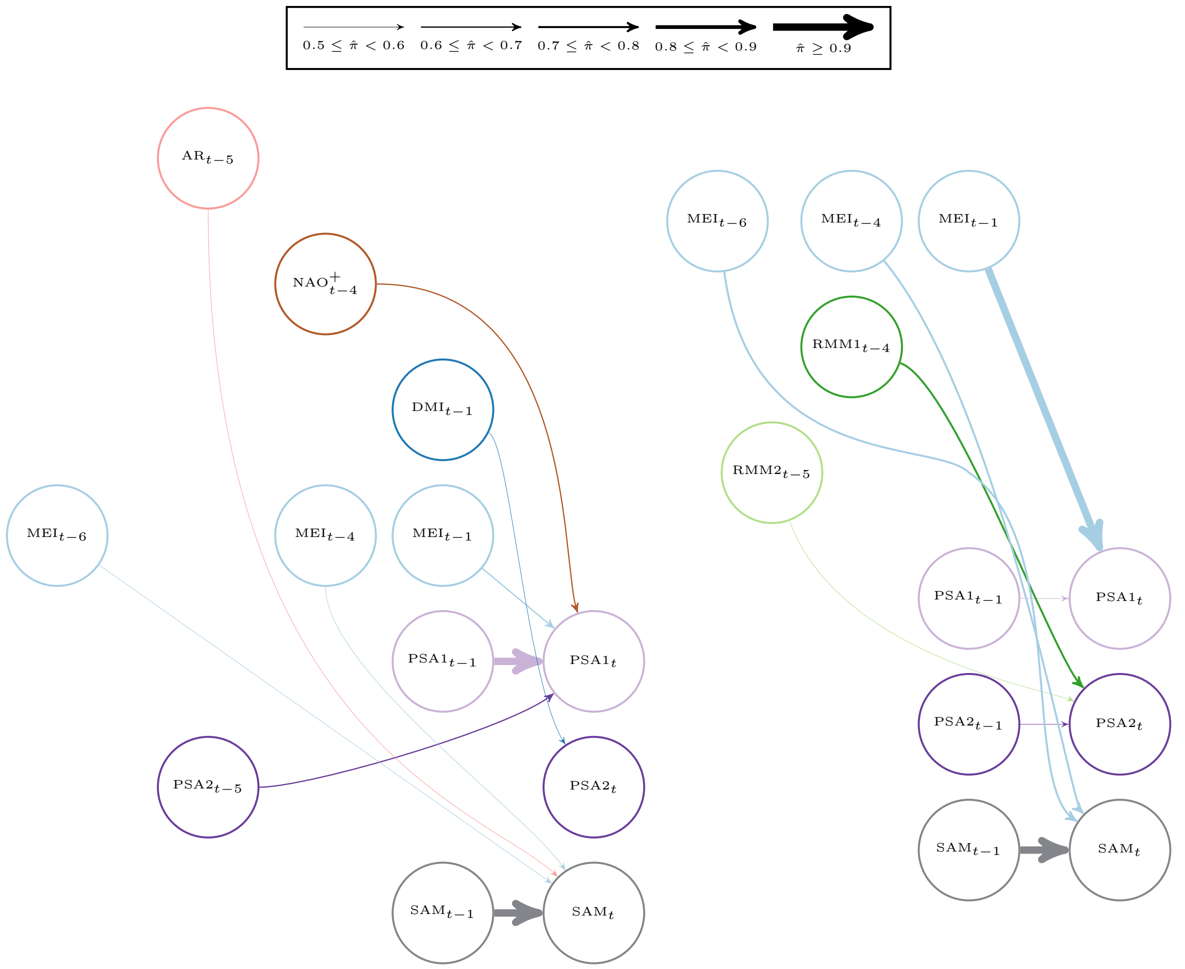


Figure S13. Subgraphs corresponding to the fitted parent sets of the tropical indices in JRA-55 based on full-year data for $a_\tau = 0.5$, $b_\tau = 10$, and $\nu^2 \approx 2$. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.



(a) NNR1 NH extratropical, all seasons (b) JRA-55 NH extratropical, all seasons

Figure S14. Subgraphs corresponding to the fitted parent sets of the NH extratropical indices in (a) NNR1 and (b) JRA-55 for $a_\tau = 0.5$, $b_\tau = 10$, and $\nu^2 \approx 2$. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.



(a) NNR1 SH extratropical, all seasons

(b) JRA-55 SH extratropical, all seasons

Figure S15. Subgraphs corresponding to the fitted parent sets of the SH extratropical indices in (a) NNR1 and (b) JRA-55 for $a_\tau = 0.5$, $b_\tau = 10$, and $\nu^2 \approx 2$. All edges with an estimated posterior probability $\hat{\pi}$ greater than 0.5 are shown.

Table S1. MAP parent sets for monthly tropical teleconnection indices across all seasons for NNR1 and JRA-55 for fits with $a_\tau = 1.5$, $b_\tau = 20$, and $\nu^2 = 3$, showing the estimated posterior probability $\hat{\pi}$ of the edge, the mean parameter value $\hat{\beta}$ conditional on the MAP structure, and the 95% posterior HDI. Dashes indicate a node that is not in the MAP parent set for a given reanalysis.

	JRA-55			NNR1		
Parent node	$\hat{\pi}$	$\hat{\beta}$	95% HDI	$\hat{\pi}$	$\hat{\beta}$	95% HDI
DMI_t						
DMI_{t-1}	1.00	0.63	$(0.56, 0.69)$	1.00	0.74	$(0.68, 0.79)$
SAM_{t-2}	0.59	0.09	$(0.03, 0.16)$	0.33	–	–
MEI_{t-1}	0.22	–	–	0.80	0.13	$(0.06, 0.20)$
MEI_{t-5}	0.08	–	–	0.65	−0.14	$(-0.21, -0.07)$
AR_{t-1}	0.04	–	–	0.94	0.10	$(0.05, 0.16)$
MEI_t						
MEI_{t-1}	1.00	1.45	$(1.37, 1.54)$	1.00	1.43	$(1.34, 1.51)$
MEI_{t-2}	1.00	−0.85	$(-1.00, -0.70)$	1.00	−0.73	$(-0.87, -0.59)$
MEI_{t-3}	1.00	0.65	$(0.49, 0.80)$	1.00	0.38	$(0.24, 0.52)$
MEI_{t-4}	1.00	−0.48	$(-0.63, -0.31)$	0.98	−0.15	$(-0.23, -0.07)$
MEI_{t-5}	0.65	0.28	$(0.13, 0.42)$	0.17	–	–
MEI_{t-6}	0.57	−0.14	$(-0.22, -0.06)$	0.10	–	–
NAO^{+}_{t-2}	0.38	0.04	$(0.01, 0.06)$	0.17	–	–
AO_{t-2}	0.06	–	–	0.46	0.05	$(0.02, 0.07)$
AR_{t-3}	0.05	–	–	0.66	−0.04	$(-0.07, -0.02)$
$\mathrm{RMM2}_{t-3}$	0.34	–	–	0.41	0.04	$(0.01, 0.06)$
$\mathrm{RMM1}_t$						
MEI_{t-3}	0.47	−0.37	$(-0.49, -0.24)$	0.37	–	–
MEI_{t-6}	0.61	0.24	$(0.13, 0.36)$	0.33	–	–
AR_{t-1}	0.54	−0.12	$(-0.20, -0.04)$	0.24	–	–
$\mathrm{PSA1}_{t-3}$	0.38	−0.14	$(-0.22, -0.05)$	0.11	–	–
SAM_{t-3}	0.34	−0.13	$(-0.20, -0.04)$	0.09	–	–
$\mathrm{RMM1}_{t-1}$	0.40	−0.12	$(-0.19, -0.04)$	0.86	−0.14	$(-0.21, -0.06)$
$\mathrm{RMM1}_{t-3}$	0.99	−0.17	$(-0.25, -0.09)$	0.97	−0.16	$(-0.24, -0.08)$
$\mathrm{RMM1}_{t-4}$	1.00	−0.20	$(-0.27, -0.12)$	1.00	−0.19	$(-0.27, -0.11)$
$\mathrm{RMM2}_{t-1}$	0.99	−0.17	$(-0.25, -0.09)$	1.00	−0.20	$(-0.27, -0.12)$
MEI_{t-4}	0.58	–	–	0.59	−0.45	$(-0.68, -0.22)$
MEI_{t-5}	0.41	–	–	0.35	0.34	$(0.11, 0.57)$
$\mathrm{RMM2}_t$						
AO_{t-1}	0.94	0.18	$(0.10, 0.26)$	0.63	0.15	$(0.07, 0.23)$
DMI_{t-5}	0.50	−0.12	$(-0.20, -0.04)$	0.06	–	–
MEI_{t-4}	0.56	−0.15	$(-0.23, -0.07)$	0.52	–	–
AR_{t-1}	0.44	0.13	$(0.05, 0.21)$	0.93	0.16	$(0.08, 0.24)$
$\mathrm{RMM1}_{t-1}$	0.71	0.13	$(0.05, 0.21)$	0.26	0.14	$(0.06, 0.22)$
$\mathrm{RMM2}_{t-1}$	0.61	−0.13	$(-0.21, -0.05)$	0.45	–	–
$\mathrm{RMM2}_{t-3}$	1.00	−0.24	$(-0.32, -0.16)$	0.98	−0.16	$(-0.23, -0.08)$
$\mathrm{RMM2}_{t-4}$	0.94	−0.17	$(-0.25, -0.09)$	0.94	−0.16	$(-0.24, -0.09)$
MEI_{t-1}	0.10	–	–	0.22	−0.18	$(-0.26, -0.10)$
AR_{t-6}	0.03	–	–	0.33	−0.11	$(-0.19, -0.04)$
$\mathrm{PSA1}_{t-3}$	0.05	–	–	0.59	−0.13	$(-0.22, -0.05)$
$\mathrm{RMM1}_{t-5}$	0.03	–	–	0.20	0.11	$(0.03, 0.19)$
$\mathrm{RMM2}_{t-2}$	0.02	–	–	0.16	0.13	$(0.05, 0.20)$

Table S2. MAP parent sets for monthly SH extratropical teleconnection indices across all seasons for NNR1 and JRA-55 for fits with $a_\tau = 1.5$, $b_\tau = 20$, and $\nu^2 = 3$, showing the estimated posterior probability $\hat{\pi}$ of the edge, the mean parameter value $\hat{\beta}$ conditional on the MAP structure, and the 95% posterior HDI. Dashes indicate a node that is not in the MAP parent set for a given reanalysis.

	JRA-55			NNR1		
Parent node	$\hat{\pi}$	$\hat{\beta}$	95% HDI	$\hat{\pi}$	$\hat{\beta}$	95% HDI
$\mathrm{PSA1}_t$						
MEI_{t-1}	0.94	-0.15	$(-0.23, -0.07)$	0.68	-0.13	$(-0.21, -0.05)$
NAO^+_{t-4}	0.39	-0.11	$(-0.20, -0.04)$	0.60	-0.13	$(-0.21, -0.05)$
$\mathrm{PSA1}_{t-1}$	0.49	0.12	$(0.04, 0.21)$	0.96	0.16	$(0.08, 0.24)$
$\mathrm{PSA2}_{t-5}$	0.17	–	–	0.58	-0.12	$(-0.21, -0.04)$
$\mathrm{PSA2}_t$						
$\mathrm{PSA1}_{t-4}$	0.31	0.12	$(0.03, 0.20)$	0.30	–	–
$\mathrm{PSA2}_{t-1}$	0.49	0.12	$(0.04, 0.21)$	0.34	–	–
$\mathrm{RMM1}_{t-4}$	0.73	0.14	$(0.05, 0.21)$	0.43	–	–
$\mathrm{RMM2}_{t-5}$	0.47	0.12	$(0.04, 0.20)$	0.41	–	–
DMI_{t-1}	0.07	–	–	0.54	-0.14	$(-0.23, -0.06)$
SAM_t						
MEI_{t-4}	0.75	-0.30	$(-0.45, -0.16)$	0.51	–	–
MEI_{t-6}	0.76	0.27	$(0.13, 0.42)$	0.54	–	–
SAM_{t-1}	1.00	0.29	$(0.21, 0.37)$	1.00	0.32	$(0.24, 0.40)$