Dynamics Bayesian Networks for Climate Model Evaluation

ACCESS-NRI ML Workshop

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

1 CSIRO Environment, Hobart
2 South Australian Health and Medical Research Institute (SAHMRI)
3 CSIRO Environment, Aspendale

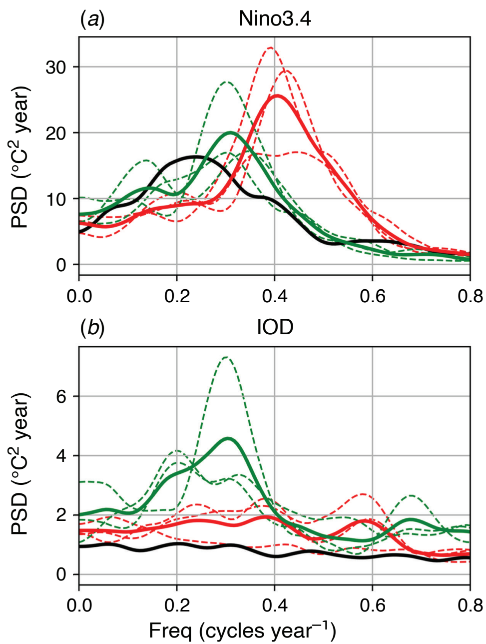
References:

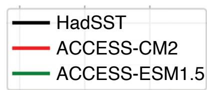
O'Kane, T. J. , Harries, D., & Collier, M.A. (2023). Dynamic Bayesian networks for evaluation of Granger causal relationships in climate models. Journal of Advances in Modeling Earth Systems, (in prep.)

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

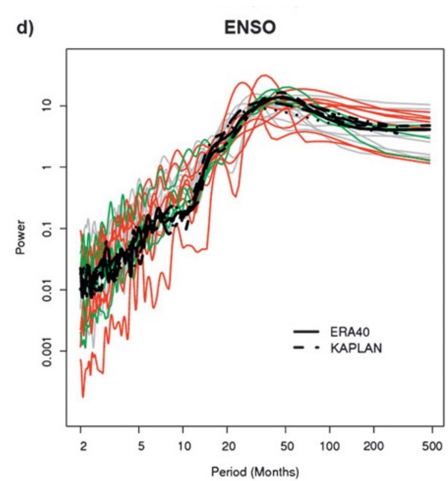
Motivation: Diagnosing climate model biases is a hugely complicated task.
Aim: Develop a mathematical, systematic approach to identify biases in the temporal behaviours of the climate modes, and then to attribute cause and effect of those biases including uncertainty estimation.

HadISST; broad peak between 3-7 years
ACCESS-CM2: 2.5 years;
ACCESS-ESM1.5: 3 years





Rashid HA *et al.* (2022)
Journal of Southern Hemisphere Earth Systems Science
72(2), 73–92. doi:10.1071/ES21028



Stoner et al. (2009) J. Clim. **22**, 438-4372, DOI: 10.1175/2009JCLI2577.1 (peaks 2-7 years)

OVERVIEW

Approach: Using the methods of Bayesian inference for causal discovery to robustly identify biases in CMIP models in terms of their ability to reproduce the observed teleconnections between the major internal modes of variability

Method: model data within a framework of a structural causal model to

1. Robust identification of probability (uncertainty) that a causal relationship exists between a given climate mode at the present time and any other set of time lagged climate indices i.e., posterior distribution
2. Robust identification of the strength of the Granger causal teleconnection i.e., lagged correlation in terms of the posterior mean
3. Compare structural causal models (RJMCMC/MC^3) sampled from CMIP5 models to those from reanalyses

General approach

Fitting a homogeneous DBN to an observed timeseries $D = x_1, \dots, x_t$, where x_t denotes the values of random variables $X_t = (X_t^1, \dots, X_t^n)^T$ at time t, requires learning the structure of the graph and the values of the corresponding parameters θ.

The two-step process is therefore:

1. The structure learning stage, the structure of the graph G is sought, independent of specific values of the parameters i.e., directly sampling the posterior $P(G|D)$.

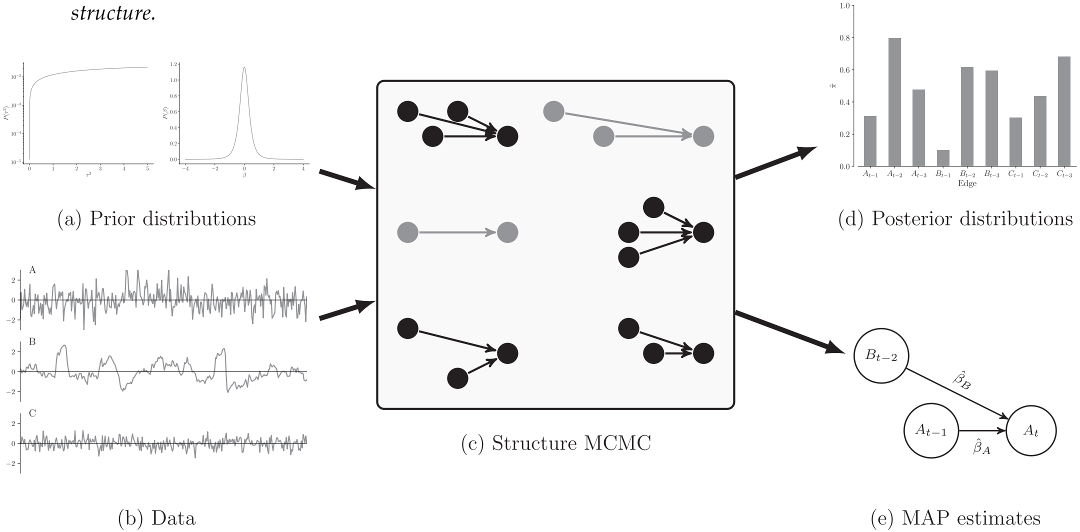
- $P(G)$ is the estimate of the (prior) probability of the hypothesis before the data D is observed.
- $P(G|D)$ is the probability of the hypothesis G given the observed evidence D.

2. Apply a Bayesian score-based approach such that the graph G is estimated based on maximizing a suitable score function, in this case the marginal likelihood P(D|G)

- The marginal likelihood is the probability of generating the observed sample from a prior distribution (likelihood function that has been integrated over parameter space) i.e., $\mathrm{P}(D|G) = \int d\theta P(D|G,\theta)P(\theta|G)$
- $P(D|G,\theta) = \prod_{t=1}^{T}\prod_{i=1}^{n} P\left(X_t^i | pa_G\left(X_t^i\right), \theta_i\right)$ is the likelihood under the model.
- $pa_G\left(X_t^i\right) = \left(X^i | G \text{ contains and edge from } X^i \text{ to } X^j\right)$
- $P(\theta|G)$ denotes a set of priors for the full set of node PDF parameters conditional on the structure of the graph.
- *Rather than find a single optimal model we sample from the full posterior distribution of possible graphs P(G|D)*
- *G provides a graphical representation of the joint pdf* $P(X^1, \dots, X^N)$

Given a set of (a) prior distributions and (b) observed data (here, time series of climate indices) a sample from the posterior distribution over possible structures is generated via (c) a structure Markov Chain Monte Carlo simulation.

Subsequently, we evaluate (d) posterior distributions for the presence of individual edges and (e) the MAP estimates for the structure.



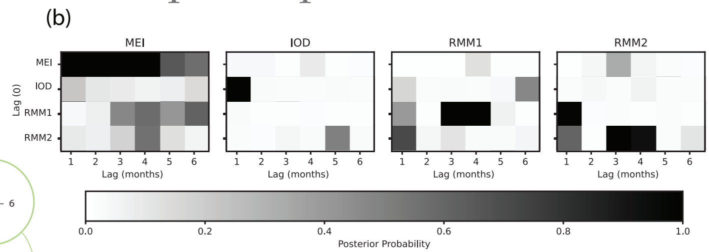
(a) Prior distributions (b) Data (c) Structure MCMC (d) Posterior distributions (e) MAP estimates

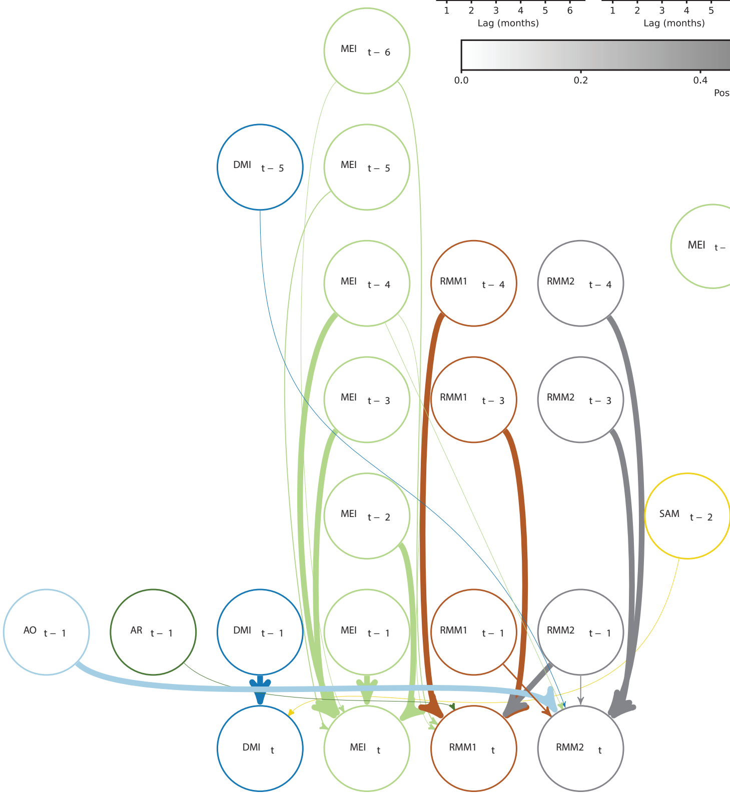
Given a sample from the model posterior distribution, by averaging over the set of possible models the posterior credibility of given features may instead be estimated in the Bayesian approach to identify edges that are well supported by the data.

- In general, reversible jump MCMC is used to sample from the joint posterior distribution P(θ,G | D) (Algorithm 1 Harries & O'Kane 2021).

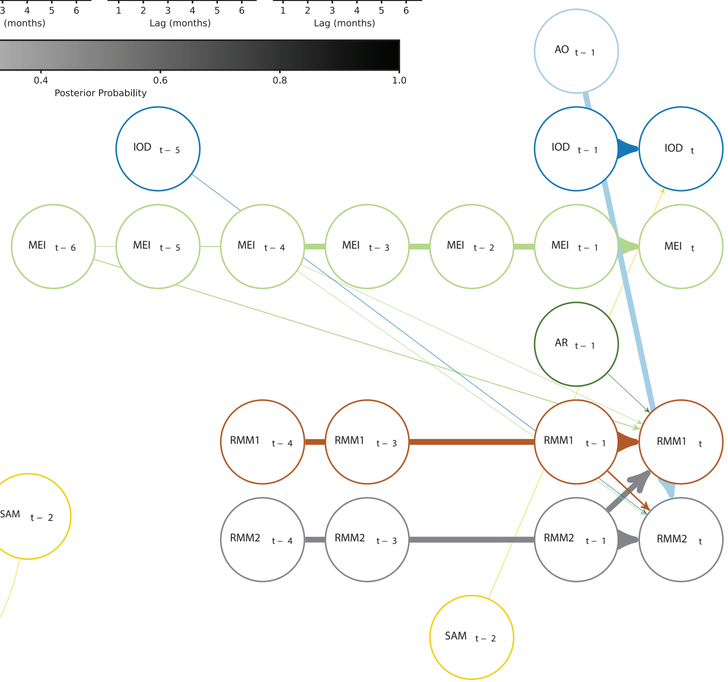
- For models where the conditional posterior distribution for all parameters admits evaluation the scheme reduces to the MC^3 scheme of Madigan et al. (1995) Algorithm 2 Harries & O'Kane 2021).

DAGs versus heatmaps: Tropical JRA55





As in Harries & O'Kane (2021)



JRA55: 'tropical', 'MEI', 'IOD', 'RMM1', 'RMM2'

$0.5 \le \hat{\pi} < 0.6$ $0.6 \le \hat{\pi} < 0.7$ $0.7 \le \hat{\pi} < 0.8$ $0.8 \le \hat{\pi} < 0.9$ $\hat{\pi} \ge 0.9$

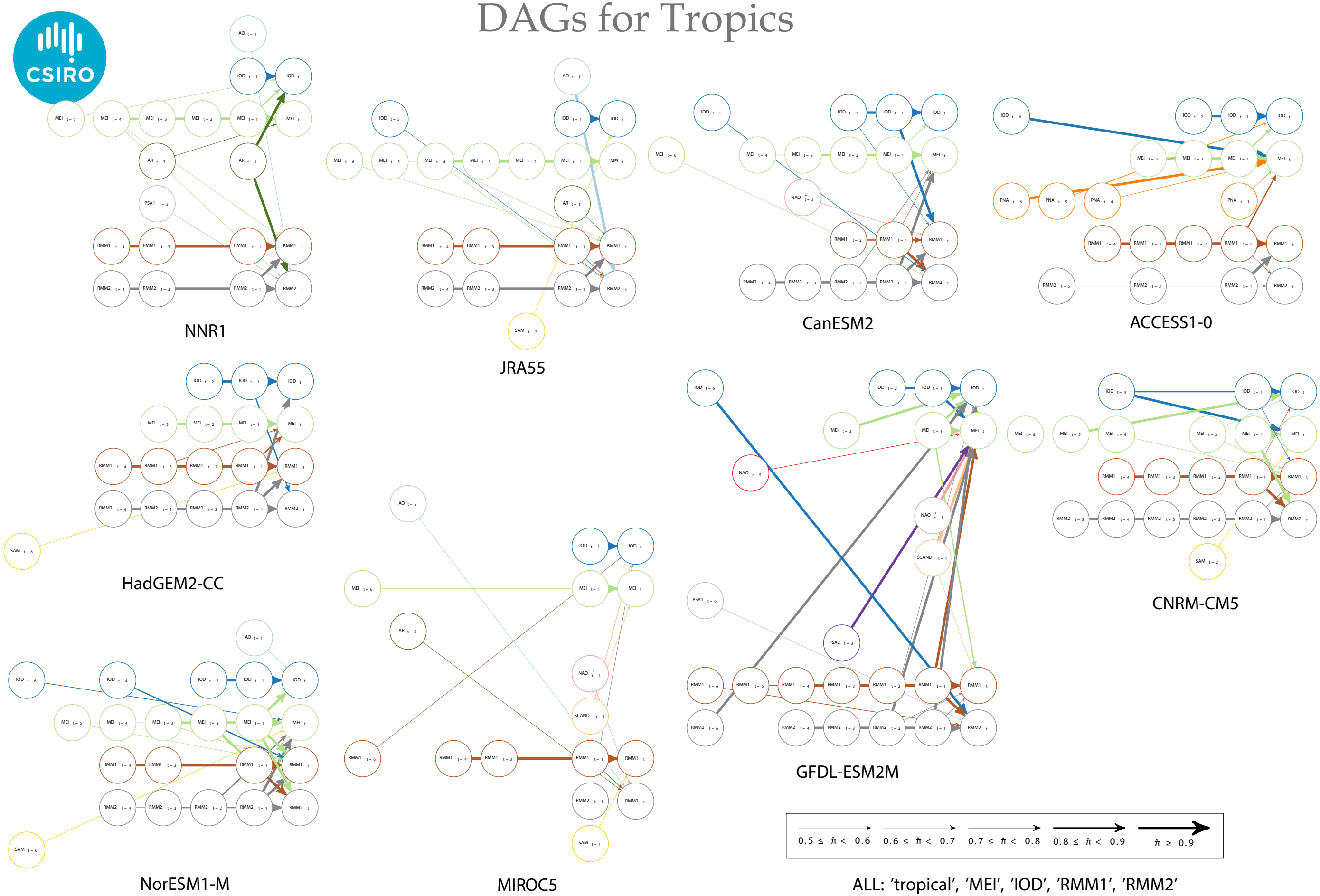
Calculation details

- Structurally modular priors for the parent sets $pa_G(Y_t^i)$, allows the prior for a graph G to be decomposed into independent priors for the parent sets $pa_G(Y_t^i)$ of each of the models n indices.
- We fix a maximum lag of τ_{max}, such that τ_{max}= 6-months.
- To sparcify the network, we set the maximum parent set size for each index to $|pa_G(Y_t^i)| \leq p_{max} = 10$.
- Hyperparameters for the priors were determined from the reanalyses and correspond to independent t_3 marginal priors for regression coefficients with 95% highest density intervals $-1 \leq \beta \leq 1$.
- Hyperparameter ranges were tested ranging from weakly informative to somewhat informative were determined to give qualitatively similar results.
- As the indices at t = 0 can be factored, for a given index, posterior samples were obtained by running 8 chains of length 1×10^7 samples (40 cycles of 250,000).
- Chain convergence was assessed by considering the homogeneity of the distribution of parent sets within chains using χ^2 and Kolmogorov-Smirnov tests (Brooks et al., 2003) for each index.
- Wasserstein distance and Kullback-Leibler divergence used to assess model performance.
- Various choices of thinning parameter were considered to determine the number of retained samples based on convergence rates. *Qualitatively our finding was that the evaluated graphs were relatively insensitive to thinning up to a factor of 1000.*
- Acceptance rates for RJMCMC were between 0.16-0.4

- 13 teleconnections (AO, SCAND, AR, NAO(+,-), PNA, PSA(1&2), MJO (RMM1&2), ENSO(MEI), IOD)
- 2 reanalyses (JRA55, NNR1)
- 7 CMIP5 models (HadGEM2-CC, CNRM-CN5, NorESM1-M, GFDL-ESM2, MIROC5, CanESM2, ACCESS1-0)
- Full year and boreal winter (DJF)

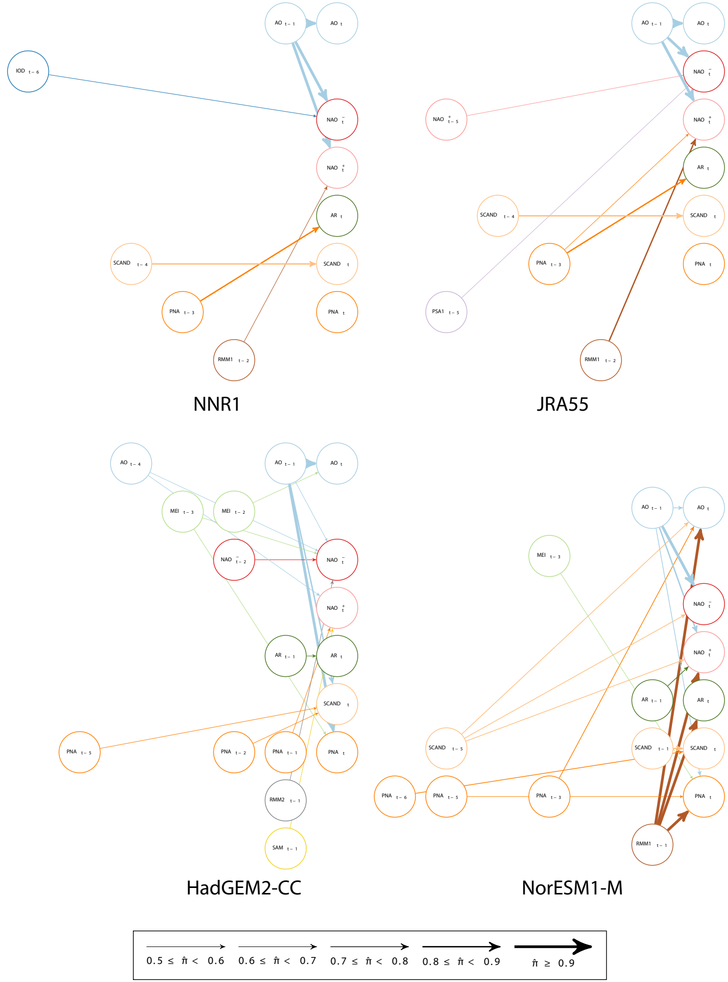
- 2,340,000,000 samples in total.

DAGs for Tropics



Thresholded ≥ 0.5

NH: DJF



DJF: 'nhtele', 'AO', 'PNA', 'NAO-', 'NAO+', 'AR', 'SCAND'

Ordering w.r.t. reference reanalysis (JRA55)

- The Wasserstein distance $W_p(P, Q)$ is the minimum amount of work required to transform one distribution into another.
- The Kullback-Leibler divergence $D_{KL}(P||Q)$ measures the information gain when the prior Q is replaced by the posterior P.

Models ordered by geographically determined Wasserstein distance						
Region →	All (global)	Tropical	NH	SH	NH-Tropical	SH-Tropical
Order ↓	JRA55	JRA55	JRA55	JRA55	JRA55	JRA55
1	NNR1	NNR1	NNR1	NNR1	NNR1	NNR1
2	HadGEM2-CC	MIROC5	MIROC5	HadGEM2-CC	HadGEM2-CC	HadGEM2-CC
3	CNRM-CM5	HadGEM2-C	CNRM-CM5	NorESM1-M	CNRM-CM5	CanESM2
4	NorESM1-M	CanESM2	HadGEM2-CC	ACCESS1-0	GFDL-ESM2M	MIROC5
5	GFDL-ESM2M	NorESM1-M	GFDL-ESM2M	CNRM-CM5	CanESM2	NorESM1-M
6	MIROC5	ACCESS1-0	ACCESS1-0	GFDL-ESM2M	ACCESS1-0	CNRM-CM5
7	CanESM2	GFDL-ESM2M	CanESM2	CanESM2	MIROC	GFDL-ESM2M
8	ACCESS1-0	CNRM-CM5	NorESM1-M	MIROC5	NorESM1-M	ACCESS1-0

Models ordered by geographically determined Kullback-Leibler divergence						
Region →	All (global)	Tropical	NH	SH	NH-Tropical	SH-Tropical
Order ↓	JRA55	JRA55	JRA55	JRA55	JRA55	JRA55
1	NNR1	NNR1	NNR1	NNR1	NNR1	NNR1
2	MIROC5	ACCESS1-0	HadGEM2-CC	ACCESS1-0	MIROC5	ACCESS1-0
3	CNRM-CM5	HadGEM2-CC	NorESM1-M	MIROC5	CNRM-CM5	CanESM2
4	ACCESS1-0	CanESM2	ACCESS1-0	CanESM2	ACCESS1-0	HadGEM2-C
5	HadGEM2-CC	MIROC5	GFDL-ESM2M	CNRM-CM5	HadGEM2-CC	MIROC5
6	NorESM1-M	CNRM-CM5	MIROC5	HadGEM2-CC	NorESM1-M	CNRM-CM5
7	CanESM2	NorESM1-M	CNRM-CM5	NorESM1-M	CanESM2	NorESM1-M
8	GFDL-ESM2M	GFDL-ESM2M	CanESM2	GFDL-ESM2M	HadGEM2-CC	GFDL-ESM2M

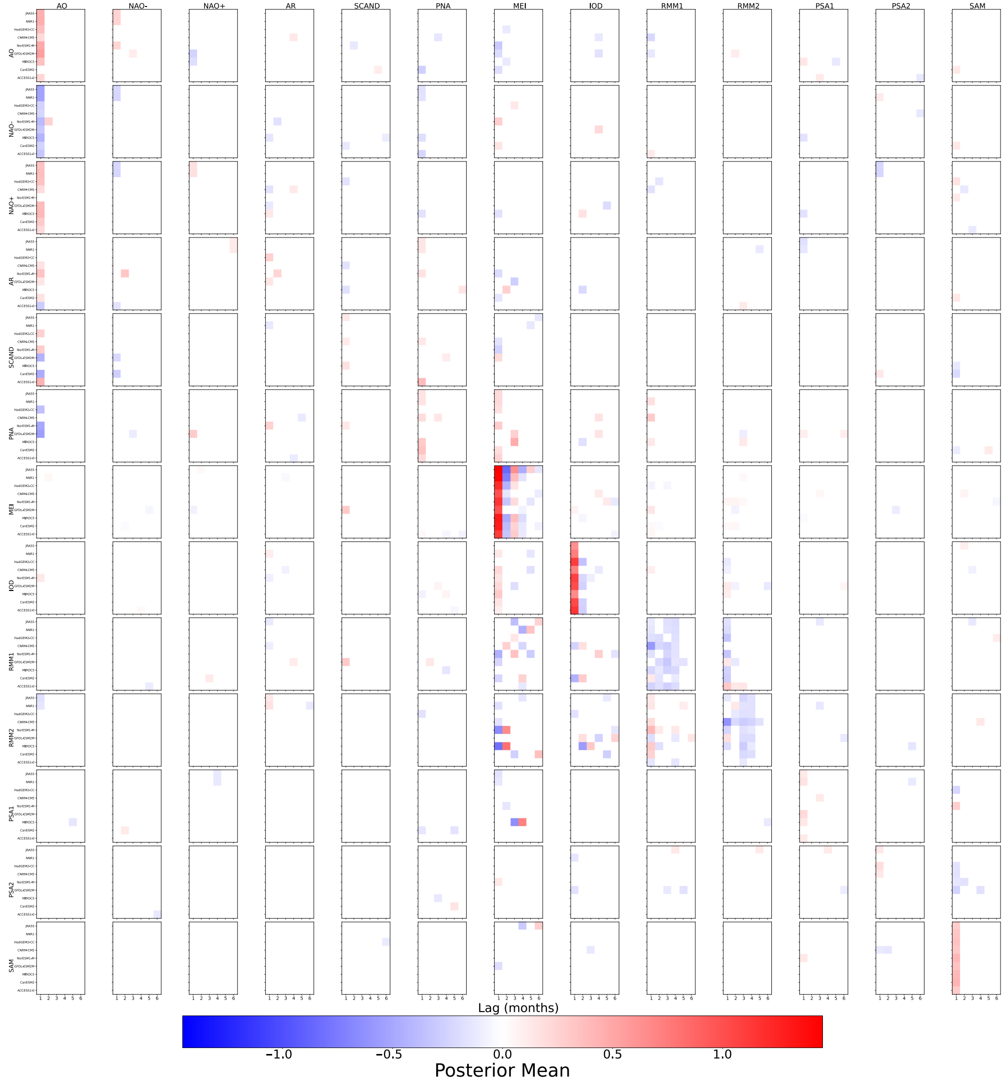
NH

Tropics

SH

Maximum a posteriori models (posterior means) increase sparsity.

Where an edge also appears in the maximum a posteriori (MAP) for the structure, the posterior 95% highest density interval (HDI) informs the choice of priors.



Key findings

- Regardless of metric, JRA55 & NNR1 reanalyses are in good agreement.
  - This is despite different models, resolutions, data assimilation methodologies (4D-VAR, EnKF), changes in observing networks and over a decade between systems becoming operational.

- For the CMIP5 models considered

1. In the tropics – where autocorrelations are large – we see the highest degree of correspondence.

2. For tropical teleconnections to the NH there is a large subset of models that agree for specific causal influences.

3. For NH influences on the tropics the models show diverse highly uncertain and weak teleconnections.

4. Tropical ⟺ SH teleconnections have generally poor correspondence to reanalysis.

5. The high latitude annular modes AO & SAM and their autocorrelations are the better simulated tropospheric modes, although their teleconnections are highly diverse across CMIP models.

Summary

- The range of spatio-temporal scales inherent in the climate system makes the general problem of Bayesian inference challenging.

- We have outlined a Bayesian framework for estimating time-homogeneous structural causal models (SCMs) based on
  1. decomposing climate data using empirical indices of the climate modes of variability
  2. learning a posterior distribution over possible structures, rather than selecting a single graph, to ascertain overall model uncertainty.
  3. identifying robust structural differences between a set of baseline estimates from observational products and climate models to establish model biases.

- Given the results here (and elsewhere), major questions arise as to the utility of GCMs to reproduce observed teleconnection behaviours and hence to inform future changes in climate teleconnections (risk).

- However, regime dependent SCMs have the potential to deepen our understanding of the future impacts of anthropogenic forcing on internal climate variability and associated climate risks and their uncertainties.

Next steps ...

6 REGIME DEPENDENT CAUSAL GRAPHS

Assume that the data can be modelled by a regime dependent (stationary) discrete-time structural causal model (SCM) i.e.,

$$\mathbf{x}_t = \hat{\mathbf{G}}_t(\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-\tau} : \theta_t) \tag{14}$$

$$\hat{\mathbf{G}}_t = [\hat{\mathbf{g}}_t^1, \ldots, \hat{\mathbf{g}}_t^N] \tag{15}$$

$$X_t^j = \mathbf{g}_t^j(\mathbf{pa}_t^j, \eta_t^j) \quad j = 1, \ldots, N \tag{16}$$

where

- η_t^j are jointly independent noise variables sampled from some stationary distribution
- $\mathbf{pa}_t^j \subset X_{t-1}, X_{t-2}, \ldots, X_{t-\tau}$ define the causal parents of X_t for some lag τ
- and we define the persistency of the respective regimes assuming that the data (time series) can be subdivided into a finite numer of K regimes i.e., $k \in 1, \ldots, K$ and that the parents **pa** and functional dependencies are stationary for on average M consecutive timesteps s.t. $K \leqslant T/M$.

Solving the inverse problem now requires finding the unknown parameters $\theta_t = [\mathbf{pa}_t, \phi_t]$ for any given data segment $\mathbf{x}_t \in \mathrm{Re}^N \ \forall t \in [0, T]$ i.e.,
$\theta_t = [\Gamma(t), \mathbf{pa}, \phi]$ where

- $\mathbf{pa}, \phi = \{\mathbf{pa}_1, \ldots, \mathbf{pa}_k; \phi_1, \ldots, \phi_k\}$
- $\Gamma(t) = [\gamma_1(t), \ldots, \gamma_k(t)]$

where $\Gamma(t) \in [0, 1]^{k \times T}$ is the probability that any particular data instance $\mathbf{x}_t^j$ resides in any of the k regimes.

We now define the cost function as:

$$L(\Gamma, \mathbf{pa}, \phi) = \sum_{t=0}^{T} \sum_{k=1}^{K} \gamma_k(t) \parallel \mathbf{x}_t - \hat{\mathbf{G}}_k(\mathbf{pa}_k, \phi_k) \parallel_F^2 \tag{17}$$

with

- $||.||_F^2$ is the squared distance function. This could be the Euclidean L2 norm or alternately based on the marginal likelihood,
- $\gamma_k(t)$ is the weight of the k regime specific network at time t
- assuming convex i.e., $\sum_{t=1}^{K} \gamma_k(t) = 1 \quad \forall t$ where $\gamma_k(t) \in [0, 1]$.

Persistency of regimes (temporal regularization) can be applied in the standard way i.e.,

$$\sum_{t=1}^{T=1} [\gamma_k(t+1) - \gamma_k(t)] \leqslant C \quad \forall k \in K. \tag{18}$$

Here $C \approxeq T/(MK)$ assumes an average regime duration of M timesteps across all K regimes but NOT that all regimes have a constant duration within the bound of having a maximum number of C transitions.

Thank You

It is then necessary to construct a Markov Chain Monte Carlo (MCMC) sampler that samples from the joint posterior $P(\theta, G|D)$ using, for example, reversible jump MCMC (RJMCMC). Here we use the linear regression model where we assume each X_t^i is assumed to be conditionally Gaussianly distributed i.e.

$$X_t^i | \mathbf{pa}(X_t^i), \tau_i^2 \sim N(\mu_t^i, \tau_i^{-2}) \tag{36a}$$

where

$$\mu_t^i = \beta_0^i + \sum_{j=1}^{p_i} \beta^i_{(k_j,\tau_j)} X^{k_j}_{t-\tau_j} \tag{36b}$$

with mean $\mu_t^i = E[X_t^i | \mathbf{pa}(X_t^i)]$ given by a linear function of the parent variables $\mathbf{pa}(X_t^i) = \{X^{k_j}_{t-\tau_j} | j = 1, \ldots, p_i\}$. The local marginal likelihoods $\Psi(D|G)$ 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 $\beta^i_{(k_j,\tau_j)}$ where

$$\tau_i^2 \sim \mathrm{Gamma}(a_\tau, b_\tau) \tag{37}$$

$$\beta^i_{(k_j,\tau_j)} | \tau_i^2, \mathbf{pa}(X_t^i) \sim N(0, \frac{\nu_i^2}{\tau_i^2}), \quad j = 1, \ldots, p_i \tag{38}$$

and a_τ, b_τ, and ν_i^2 are prior hyperparameters.