A hidden semi-Markov model for characterising regime shifts in ocean density variability

Theodoros Economou[*], Matthew B. Menary[†]

Abstract

Societally important decadal predictions of temperature and precipitation over Europe are largely affected by variability in the North Atlantic Ocean. Within this region, the Labrador Sea is of particular importance due its link between surface-driven density variability and the Atlantic Meridional Overturning Circulation (AMOC). Using physical justifications, we propose a statistical model to describe the temporal variability of ocean density in terms of salinity-driven and temperature-driven density. This is a hidden semi-Markov model that allows for either a salinity-driven or a temperature-driven ocean density regime, such that the persistence in each regime is governed probabilistically by a semi-Markov chain. The model is fitted in the Bayesian framework, and a reversible MCMC algorithm is proposed to deal with a single-regime scenario. The model is first applied to a reanalysis data set, where model checking measures are also proposed. Then it is applied to data from 43 climate models to investigate whether and how ocean density variability differs between them and also the reanalysis data. Parameter estimates relating to the mean holding time for each regime are used to establish a link between regime behaviour and the AMOC.

Keywords: Reversible jump; Bayesian; MCMC; HMM; Forward algorithm; Adaptive Metropolis.

1 Introduction

Skillful decadal predictions of changes in temperature and precipitation over Europe are valuable for society—for example, to plan adaptation/mitigation strategies and

[*]Department of Mathematics, University of Exeter, UK
[†]LOCEAN/IPSL, Sorbonne Universités (UPMC)-CNRS-IRD-MNHN, Paris, France

maximise economic growth (e.g. through energy trading). In recent years, such experimental predictions have begun to be made, but there remain significant scientific challenges. For example, on decadal timescales, low frequency variability in the North Atlantic ocean (*e.g.* the Atlantic Meridional Overturning Circulation, AMOC) becomes an increasingly important component of any skilful prediction (Collins, 2002). This low frequency variability has often been linked to the Labrador Sea region, and variability in seawater density therein (Ba et al., 2014). Understanding the nature of seawater density changes in this region could thus be valuable in improving the skill of predictions.

The relatively short time space of instrumental observations, means that it is not clear whether the interannual variability in density in this region is driven by salinity or temperature—with implications for the processes involved and subsequent evolution of the region (Menary et al., 2015). It is therefore necessary to rely on coupled general circulation climate models, in order to understand the multiannual density variability and its drivers. More specifically, recent analysis suggests that the real world Labrador Sea may have recently transitioned from a salinity to a temperature dominated regime (Menary et al., 2016). In order to estimate whether any recent regime shift is either a temporary or a longer-term (multi-decadal) change, it makes sense to investigate the nature/transience of Labrador Sea density-drivers throughout a number of different climate models.

The work in this paper is motivated by two research questions. The first: are water density changes in the Labrador Sea driven by sea temperature or salinity changes and is the driver of density variability stationary in time, or are there regime shifts between salinity-driven and temperature-driven density over time? Second: what can we learn from climate models about this potential non-stationarity? As such, there are two goals here. The first is to develop an appropriate statistical model for quantifying the temporal variability in the Labrador Sea density in terms of salinity and temperature regime shifts, and apply it to available observations. The second is to apply this model to data from 43 free-running climate models, and use the results to investigate whether climate models are able to simulate regime shifts and thus better understand how long these shifts can last for.

The paper is structured as follows. Section 2 presents the background in terms of physical understanding of ocean density variability, which leads nicely to the form of statistical models to be used. Section 3 then introduces hidden Markov and semi-Markov models, presenting the argument for why these are appropriate for the questions posed here, and then proceeding to describe the proposed modelling framework. Subsequently, Section 4 presents the implementation of the proposed model to ocean density observations, but also to density data from 43 climate models. Finally Section 5 gives a short summary and a discussion.

2 Background

The importance of understanding the Labrador Sea density variability in terms of temperature and salinity variability has been highlighted in Menary et al. (2016). Seawater density ρ, can be described mathematically as a function of temperature T and salinity S such that

$$\rho = f_T(T) + f_S(S) + f_{S,T}(S,T).$$

This equation is non-linear in temperature and salinity over the full, observed temperature/salinity parameter space. However, for small perturbations (i.e. of the order of interannual variability averaged over a large enough volume) around a given mean Temperature/Salinity state ($\bar{T}$, $\bar{S}$), the equation can be considered approximately linear. Here, we focus on the top 500m of the North Atlantic Labrador Sea (45-60°W, 55-65°N), for which a linear approximation of T and S explains more than 99% of the variance in most climate models (Menary et al., 2015), thus:

$$\rho \approx f_T(T,\bar{S}) + f_S(\bar{T},S) + f(\bar{T},\bar{S}). \tag{1}$$

where $\rho_0 = f(\bar{T},\bar{S})$ is a constant and is the density about which these perturbations are occuring. The approximation we use follows Delworth et al. (1993) and Menary et al. (2015) and decomposes the density (ρ) into components due solely to temperature variations (denoted ρ_T) and components due solely to salinity variations (ρ_S):

$$\begin{aligned} \rho_T &= \rho(\overline{T}+T',\overline{S}) \\ \rho_S &= \rho(\overline{T},\overline{S}+S') \end{aligned}$$

where $\overline{X}$ and X' denote temporal mean and annual anomalies (about that mean) respectively. Given that ρ_0 is time invariant and that our focus is on variability, we further consider just temporal anomalies (deviations about the mean) by subtracting ρ_0 from (1), such that

$$\begin{aligned} \rho' &\approx f_T(T,\bar{S})' + f_S(\bar{T},S)' \\ &= \rho'_T + \rho'_S. \end{aligned} \tag{2}$$

Finally, we take advantage of observed physical behaviour in the region of interest. That is, the signs of the anomalous density changes due to either temperature (ρ'_T) or salinity (ρ'_S) generally oppose one another on interannual timescales due to the oceanic anomalies generally being simultaneously either both warm and saline or both cold and fresh (Yashayaev and Loder, 2017). As such, to first order, the density anomaly (ρ'_T or ρ'_S) that has the same sign as ρ' can be said to be driving ρ'. (For clarity of exposition, we will henceforth omit the dash symbol which simply denotes anomalies.) This in conjunction with equation (2) results in two possible equations

for ρ:

$$\rho = \beta_S \rho_S + \epsilon_S \qquad \text{(salinity driven density)} \tag{3}$$

$$\rho = \beta_T \rho_T + \epsilon_T \qquad \text{(temperature driven density)} \tag{4}$$

where the uncertainty ϵ_T (ϵ_T) is partly due to concomitant variations in temperature (salinity) that are (by definition) not captured by ρ_S (ρ_T) and the scaling factor β_S (β_T) captures the density compensation (with a smaller scaling factor implying more density compensation). It is worth recalling that in this formulation, ρ, ρ_S and ρ_T are anomalies, so that their mean is zero and they are interpreted as deviations about a common mean.

2.1 Density regimes

Based on ocean reanalyses, there is evidence in observed data pointing to the fact that density changes over time are indeed described by two alternating regimes on a decadal time scale (Menary et al., 2016). One is a salinity driven regime (3) where density changes over time are mostly explained by changes in salinity, while the other is a temperature driven regime (4) where density is primarily driven by temperature. As such, we seek a modelling framework here that can describe and estimate this alternating regime change in time—in a probabilistic manner—from data sets of $\rho(t)$, $\rho_S(t)$ and $\rho_T(t)$.

3 Modelling framework

3.1 Hidden Markov and semi-Markov models

A natural modelling framework for describing underlying regime changes in the distribution of a random variable, is the hidden Markov model or HMM. This latent structure model is one where the probability distribution of the modelled quantity is assumed dependent upon the states of an unobserved discrete time Markov chain. At any given time point (where time is defined by discrete equidistant points), the hidden chain will be in a particular state (from which regimes are defined), and in each state the data generating mechanism can be different. HMMs were first introduced in the context of speech recognition and have since found use in a plethora of different applications such as environmental (Hughes et al., 1999; Bellone et al., 2000), medical (Jouyaux et al., 2000; Kozumi, 2000; Hu and Gruttola, 2007) and financial (Rydén et al., 1998).

A potential limitation in HMMs is that the state holding times are implicitly geometrically distributed, meaning that in some applications the temporal persistence of some regimes cannot be adequately captured (e.g. Guedon (2003); Tokdar et al. (2010)). A natural extension of HMMs are hidden semi-Markov models (HSMMs), where state holding times are explicitly defined. HSMMs are in general more compu-

tationally intensive than HMMs, however they have still found use in environmental applications (Sansom and Thomson, 2001) but also elsewhere—see Yu (2010) for a detailed list.

In what follows, we present a Bayesian hidden semi-Markov modelling framework for describing ocean density temperature/salinity regime changes, based on HSMMs. We employ reversible jump MCMC to allow for the special cases where the density variability is driven solely by either temperature or salinity.

3.2 Model formulation

Define ocean density by $\rho(t)$ where the time step $t = 1, \ldots, n$ is yearly. Also define by $\rho_S(t)$ and $\rho_T(t)$ the density where respectively temperature T or salinity S are being held constant. Given that under the assumptions in Section 2 all three variables have the same mean, we can mean-centre the data so that all three variables have mean zero. This has the added benefit of avoiding issues with interpretation of intercepts in the models below. As described in Section 2, a potential model for describing the situation where density is being solely driven by salinity changes across all time is:

$$\begin{aligned} \rho(t) &= \beta_S \rho_S(t) + \epsilon_S(t) && (5) \\ \epsilon_S(t) &\sim N(0, \sigma_S^2) && (6) \end{aligned}$$

and we denote this model by M_S. The equivalent model, M_T for solely temperature driven density is:

$$\begin{aligned} \rho(t) &= \beta_T \rho_T(t) + \epsilon_T(t) && (7) \\ \epsilon_T(t) &\sim N(0, \sigma_T^2). && (8) \end{aligned}$$

Both of these models are extreme in that they assume no regime shifts, whereas the hypothesis here is that at any point in time, density will either be in a salinity regime or a temperature regime. We therefore consider a third model, M_{ST} where a regime switching mechanism is described by a latent semi-Markov chain $C(t)$ with two states: $C(t) \in \{S, T\}$. The model is given by:

$$\begin{aligned} \rho(t)|C(t) &= \beta_{C(t)} \rho_{C(t)}(t) + \epsilon_{C(t)}(t) && (9) \\ \epsilon_{C(t)}(t) &\sim N(0, \sigma_{C(t)}^2). && (10) \end{aligned}$$

This is a model that jumps between M_S and M_T, as depicted in Figure 1. The semi-Markov chain is defined by two holding time distributions $h(\tau; \phi_S)$ and $h(\tau; \phi_T)$, where $\tau = 1, 2, \ldots$ is the holding time random variable and (ϕ_S, ϕ_T) are associated holding time parameters for each regime. If both these distributions are Geometric, then this is an HMM where the scalar parameters $0 < \phi_S, \phi_T < 1$ define the 2×2 transition matrix. Notice that $\tau \neq 0$ so that self-transitions are not allowed, as this would conflict with the very definition of holding times between well-defined regimes (Economou et al., 2014). This implies that neither M_S nor M_T are special cases of M_{ST}, a point we return to later. The definition of the latent chain is completed by

an initial state distribution $\boldsymbol{\pi} = (\pi_S, \pi_T)$, the probability that in the first time step $t = 1$ the state is either $C(1) = S$ or $C(1) = T$.

Note that conditional on the covariates $\rho_S(t)$ and $\rho_T(t)$, models M_S and M_T assume independence in $\rho(t)$, although marginally we would expect $\rho(t)$ to inherit temporal dependence from the covariates. On the other hand, model M_{ST} directly induces temporal structure in $\rho(t)$ since density values in each regime will be more similar to each other.

3.3 Likelihood

To define the likelihood of model M_{ST}, it is instructive to start by thinking of the likelihood of a semi-Markov chain observed in some time interval $[t = 1, t = n]$. A particular realisation of such a chain is given in the bottom half of Figure 1, where it starts in state S which holds for $\tau_1 = 6$ time steps, then switches to state T which holds for $\tau_2 = 5$ time steps and so on. Given this chain will ultimately be assumed latent, the holding time of the last state held is assumed right censored so that $\tau_K \geq 4$. The data for this chain are then the holding times $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_K)$ and the regime sequence $\boldsymbol{r} = (r_1, \ldots, r_K)$ where $r_k \in \{S, T\}$. The likelihood is defined as

$$L_{SMC}\left(\boldsymbol{\tau}, \boldsymbol{r}; \boldsymbol{\phi}, \boldsymbol{\pi}\right) = \pi_{r_1} \prod_{k=1}^{K} h_{r_k}(\tau_k; \phi_{r_1}) \times H_{r_K}(\tau_K; \phi_{r_K}) \tag{11}$$

where $\boldsymbol{\phi} = (\phi_S, \phi_T)$ and $H_{r_K}(\tau_K; \phi_{r_K}) = \Pr(\tau > \tau_K; \phi_{r_k}) = \int_{\tau_K}^{\infty} h_{r_K}(u; \phi_{r_K}) du$.

Now let's consider the conditional likelihood of model M_{ST}, given the observed semi-Markov chain during a particular regime. While in regime $r_k \in \{S, T\}$ which lasts for a time period of τ_k time steps, this is given by:

$$L_{r_k}\left(\boldsymbol{\rho}(\tau_k), \boldsymbol{\rho}_{r_k}(\tau_k); \beta_{r_k}, \sigma^2_{r_k} \mid r_k, \tau_k\right) = \prod_{t \in \tau_k} f\left(\rho(t), \rho_{r_k}(t); \beta_{r_k}, \sigma^2_{r_k}\right) \tag{12}$$

where $f\left(\rho(t), \rho_{r_k}(t); \beta_{r_k}, \sigma^2_{r_k}\right)$ is the Gaussian probability density function implied by (9)-(10) depending on the the regime r_k. Also, $\boldsymbol{\rho}(\tau_k)$ and $\boldsymbol{\rho}_{r_k}(\tau_k)$ are vectors representing the data in time interval τ_k. Considering now the whole time interval $[t = 1, t = n]$, the joint likelihood of the data and the observed chain is given by:

$$\begin{aligned} L_{SMM}\left(\boldsymbol{\rho}, \boldsymbol{\rho}_S, \boldsymbol{\rho}_T, \boldsymbol{r}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) &= \prod_{k=1}^{K} h_{r_k}(\tau_k; \phi_{r_k}) L_{r_k}\left(\boldsymbol{\rho}(\tau_k), \boldsymbol{\rho}_{r_k}(\tau_k); \beta_{r_k}, \sigma^2_{r_k} \mid r_k, \tau_k\right) && (13) \\ &\times \pi_{r_1} H_{r_K}(\tau_K; \phi_{r_K}) && (14) \end{aligned}$$

obtained by combining (11) and (12) and where $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\pi}, \beta_S, \beta_T, \sigma^2_S, \sigma^2_T)$ are all the parameters of model M_{ST}. Recall however that the chain is latent, so to obtain the marginal likelihood of the data one needs to integrate it out—something which involves a computationally intensive combinatoric sum over all possible regime

sequences and holding times:

$$L_{HSMM}\left(\boldsymbol{\rho}, \boldsymbol{\rho}_{r_k}; \boldsymbol{\theta}\right) = \sum_{\tau_1+\cdots+\tau_K=n} \sum_{r_1=S}^{T} \cdots \sum_{r_K=S}^{T} L_{SMM}\left(\boldsymbol{\rho}, \boldsymbol{\rho}_S, \boldsymbol{\rho}_T, \boldsymbol{r}, \boldsymbol{\tau}; \boldsymbol{\theta}\right) \tag{15}$$

where n is the length of the observation window in discrete time steps.

3.3.1 Forward and backward algorithm

For a given value of $\boldsymbol{\theta}$, Economou et al. (2014) provide an efficient algorithm for computing the sum in (15) and use this to fit HSMMs in a Bayesian setting using MCMC—and here we adopt this algorithm to fit model M_{ST} as a Bayesian model. This is the so called forward algorithm (in HMM jargon), which sequentially computes the probability distribution $p(C(t)|\boldsymbol{\rho}_{1:t}, \boldsymbol{\theta})$ of the latent states at each time step t given the data up to t, from which the likelihood is computed as a by-product.

For the purpose of estimating the latent state sequence, a backward algorithm can be utilised after model fitting (Economou et al., 2014), to compute the probability distribution $p(C(t)|\boldsymbol{\rho}, \boldsymbol{\theta})$ of the latent states at each time step t given all the data (not just the data up to t). Monte Carlo can then be used to integrate out $\boldsymbol{\theta}$ to obtain the predictive distribution $p(C(t)|\boldsymbol{\rho})$ of the latent states at each time step. One can then simulate from this distribution or set a probability threshold such as 0.5, to decide on the most likely state at each t.

Note, however, that this method uses the marginal distribution of each state given the data and as such will not necessarily produce the most likely state sequence, which would be obtained by maximizing $p(\boldsymbol{C}|\boldsymbol{\rho})$ with respect to $\boldsymbol{C} = \{C(t)\}$. The most likely state sequence is in many applications of primary interest, and this is certainly the case here where focus lies in the decadal variability of the regimes. Conventionally, the Viterbi algorithm (Viterbi, 1967) has been employed to compute the most likely state sequence for HMMs by maximizing $p(\boldsymbol{C}|\boldsymbol{\rho}, \boldsymbol{\theta})$, using recursive algorithms similar to the forward and backward. However, this is conditional upon the parameters $\boldsymbol{\theta}$, and in a Bayesian setting one should really integrate out $\boldsymbol{\theta}$ to quantify estimation uncertainty. Using Monte Carlo is possible, although this will not maintain the Markovian (and indeed semi-Markovian) structure of the chain. For HMMs implemented using MCMC, Scott (2002) proposed a simulation approach for approximating the most likely state sequence from $p(\boldsymbol{C}|\boldsymbol{\rho})$. This is based on viewing the most likely state sequence as a function of $\boldsymbol{\theta}^{(j)}$, where j is a particular MCMC iteration. For each $\boldsymbol{\theta}^{(j)}$ one can use the Viterbi algorithm and obtain the most likely state sequence $\boldsymbol{C}^{(j)}$, and then approximate the most likely sequence as the $\boldsymbol{C}^{(j)}$ that happens most often.

To employ this here, we would first need to adapt the Viterbi algorithm to the HSMM. Fortunately, it is fairly straightforward to modify the forward-backward algorithm described in Economou et al. (2014) in order to employ the (HMM) Viterbi algorithm described in Section 3.2 of Scott (2002), for HSMMs. This is described in more detail in the Appendix.

3.4 Model specification

Here we complete the model specification by specifying the holding time distributions and the prior distribution for $\boldsymbol{\theta}$. We follow the example in Economou et al. (2014) and assume the holding time distribution for each state is a zero-truncated Poisson distribution so that $h(\tau; \phi_{r_k}) = \phi_{r_k}^{\tau} e^{-\phi_{r_k}} / \left(\tau!(e^{\phi_{r_k}} - 1)\right)$ with mean holding time $\phi_{r_k} / \left(1 - e^{-\phi_{r_k}}\right)$. The implications of this modelling choice are checked later on, using out-of-sample prediction. For the holding time parameters ϕ_S and ϕ_T we assume independent exponentially distributed priors with rate parameter 0.1 (mean 10). This is a weakly informative prior with mean 10 and variance 100, implying an average holding time of $10/\left(1 - e^{-10}\right) \approx 10$ for each state. This is consistent with the application to ocean density where the regime changes are expected to vary on a decadal time scale. Furthermore, we assume this prior to be truncated on $(0, n)$: the lower bound to adhere with the range of $\phi_S, \phi_T > 0$ (disallow self-transitions to ensure identifiability of holding times and thus interpretation of their probability distribution) while the upper bound (the length of observation period) to aid in model identification. This upper bound reflects the fact that model M_{ST} captures a regime changing scenario, while single regime scenarios are meant to be captured by models M_S and M_T.

For the coefficients β_S and β_T, we assume a Gaussian distribution with zero mean and standard deviation $\sqrt{2}$. Although informative, this is to ensure that the response $\rho(t)$ attains plausible values of ocean density (see equations (5) and (7)). For the standard deviations σ_S and σ_T we consider an exponential prior with rate 10 (mean 0.1, variance 0.01). This was chosen to reflect the small variability in observed time series of ocean density where standard deviation estimates were at least one order of magnitude smaller than 0.1. Finally, for the initial state distribution parameter π_S we assume a uniform distribution on $(0, 1)$, noting that $\pi_T = 1 - \pi_S$.

3.5 Model implementation

There are seven parameters to estimate for model M_{ST} so we therefore consider obtaining samples from the joint posterior distribution using MCMC, and in particular the Metropolis-Hastings (MH) algorithm. More specifically, we consider a robust adaptive MH algorithm proposed by Vihola (2012) and implemented in the R package adaptMCMC (Scheidegger, 2018). This implies that all parameters are updated simultaneously using a multivariate normal proposal distribution, which is adapted to achieve a particular acceptance rate—here this is set to 0.25. Such a proposal distribution assumes parameters on the real line, meaning that all parameters except β_S and β_T must be transformed appropriately.

For parameters σ_S and σ_T we use the transformation $u_\sigma = q_\sigma(\sigma) = \log(\sigma)$ (dropping the subscript) so that the prior on u_σ is $f(u_\sigma) = f_\sigma(e^{u_\sigma})e^{u_\sigma}$, where $f_\sigma(\cdot)$ is the exponential prior on σ. The parameter π_S is transformed using $u_\pi = q_\pi(\pi) = \log(\pi/(1-\pi))$ so that the prior distribution on u_π is $f_\pi\left(1/(1+e^{-u_\pi})\right)(2 + e^{u_\pi} + e^{-u_\pi})^{-1}$, and since $f_\pi(\cdot)$ is $U(0,1)$, this simplifies to $f(u_\pi) = (2 + e^{u_\pi} + e^{-u_\pi})^{-1}$.

Finally, for parameter $\phi_S \in (0, n)$ (and thus ϕ_T) we consider $u_\phi = q_\phi(\phi) = \log(\phi/(n-\phi))$ so that $u_\phi \in (-\infty, \infty)$. The prior on u_ϕ is then $f_\phi\left(n/(1-e^{-u_\phi})\right)\left(n/(2+e^{u_\phi}+e^{-u_\phi})\right)$ where $f_\phi(\cdot)$ is the truncated exponential prior on ϕ_S.

3.6 Model selection via reversible jump

As noted earlier, neither model M_S nor M_T are special cases of model M_{ST}, due to the fact that regime holding times are considered non-zero and have a restricted upper bound through the prior on ϕ_S and ϕ_T. Allowing holding times to be zero would result in a non-identifiable model in the parameters relating to the state with zero holding time. In fact, zero holding time are equivalent to allowing self-transitions, something which is contrary to the very definition of the HSMM in terms of holding time distributions.

As such, a way is required by which to decide whether a particular data set is better explained by either of the three models: do the data support a regime changing scenario (M_{ST}) or is it a case of solely salinity (M_S) or solely temperature (M_T) driven ocean density? Given that the models do share some of the parameters, one way to do this in the Bayesian context is to use reversible jump MCMC or RJMCMC (Green, 1995), where a prior probability distribution is placed upon which model is preferable a priori, so that the data are used to derive a posterior probability for which model is more appropriate.

The general formula for the acceptance probability of going from model M_v with parameters θ_v to proposed model M_{v^*} with parameters θ_{v^*} is given by:

$$\alpha_{v,v^*} = \frac{L(\boldsymbol{\rho}; \theta_{v^*}, M_{v^*}) p(\theta_{v^*} \mid M_{v^*}) p(M_{v^*}) p(M_{v^*} \mid M_v) J_{v^*,v}(u^* \mid \theta_{v^*})}{L(\boldsymbol{\rho}; \theta_v, M_v) p(\theta_v \mid M_v) p(M_v) p(M_V \mid M_{v^*}) J_{v,v^*}(u \mid \theta_v)} \left| \frac{\partial g_{v,v^*}(\theta_v, u)}{\partial (\theta_v, u)} \right| \tag{16}$$

where $L(\boldsymbol{\rho}; \theta_{v^*}, M_{v^*})$ is the likelihood of model M_{v^*} (noting that we dropped the covariates $\boldsymbol{\rho}_S$ and $\boldsymbol{\rho}_S$ for brevity), $p(\theta_{v^*} \mid M_{v^*})$ is the joint prior of its parameters, $p(M_{v^*})$ is the prior probability on model M_{v^*} and $p(M_{v^*} \mid M_v)$ is the proposal probability of going from model M_v to M_{v^*}. Here, we assume $p(M_v) = 1/3$ for all $v = S, T, ST$ and $p(M_{v^*} \mid M_v) = 1/3$ for all $v^*, v = S, T, ST$.

To ensure the parameter dimension is matched across models, auxiliary parameters u are required for RJMCMC, with associated functions $g_{v,v^*}(\theta_{v^*}, u)$ that return a vector collecting parameters across both models (some of which can be shared). As such, $J_{v,v^*}(u \mid \theta_v)$ in (16) denotes the proposal distribution of the auxiliary parameters required for moving from model M_v to proposed model M_{v^*}. Finally, the last term in (16) denotes the Jacobian determinant of $g_{v,v^*}(\theta_v, u)$.

In specific modelling frameworks such as finite mixture models, the auxiliary variables can be chosen in way so that all parameters across all models have some interpretation, e.g. mixing probabilities of two of the components in a 3-component mixture model can be defined as a sum of one of the components in a reduced 2-component model. In the context of HMMs (which can be thought of an extension of finite mixture

models), Robert et al. (2000) have utilised RJMCMC to fit a conditional Gaussian finite mixture distribution.

However, for models with covariates (such as here) where coefficients are not necessarily interpretable across models, setting up suitable auxiliary parameters is less straightforward. Instead, we propose here to fit each of the three density models separately (keeping in mind that M_S and M_T are very simple models) and then use the resulting (adapted) proposal distributions as the $J(\cdot|\cdot)$ in (16) to conduct RJMCMC.

Below we describe in detail the various moves between each of the three models. Start by denoting the parameter vector for each model: $\theta_S = (\beta_S, \sigma_S)$, $\theta_T = (\beta_T, \sigma_T)$ and $\theta_{ST} = (\beta_S, \beta_T, \sigma_S, \sigma_T, \phi_S, \phi_T, \pi_S)$ respectively for M_S, M_T and M_{ST}. Assuming an MCMC run for each of the three models, the adaptive MH discussed earlier will provide (adapted) multivariate Gaussian proposal distributions $P_S(\cdot)$, $P_T(\cdot)$, $P_{ST}(\cdot)$ for each of $u_S = (\beta_S, q_\sigma(\sigma_S))$, $u_T = (\beta_T, q_\sigma(\sigma_T))$ and

$$u_{ST} = (\beta_S, \beta_T, q_\sigma(\sigma_S), q_\sigma(\sigma_T), q_\phi(\phi_S), q_\phi(\phi_T), q_\pi(\pi_S)), \tag{17}$$

respectively.

Moving from M_S to M_T. Current parameter vector is $\theta_S = (\beta_S, \sigma_S)$ so need to propose auxiliary variable u relating to the two parameters $\theta_T = (\beta_T, \sigma_T)$ of M_T. This can be proposed from $P_T(u)$ defined above. In the notation of (16), this implies $J_{S,T}(u|\theta_S) = J_{S,T}(u) = P_T(u)$. Then do the dimension matching using

$$g_{S,T}(\theta_S, u) = (\beta_S, \sigma_S, u_1, \exp\{u_2\}) = (u_1^*, \exp\{u_2^*\}, \beta_T^*, \sigma_T^*)$$

where u_j denotes the j^{th} element of u. Then,

$$\left| \frac{\partial g_{S,T}(\theta_S, u)}{\partial (\theta_S, u)} \right| = \exp\{u_2\}$$

and the acceptance probability is

$$\alpha_{S,T} = \frac{L(\boldsymbol{\rho}; \theta_T^*, M_T) p(\theta_T^* \mid M_T) J_{T,S}(u^*)}{L(\boldsymbol{\rho}; \theta_S, M_S) p(\theta_S \mid M_S) J_{S,T}(u)} \exp\{u_2\}$$

where $J_{T,S}(\cdot)$ is the Gaussian proposal $P_T(\cdot)$ obtained from fitting model M_T.

Moving from M_T to M_S. This move is symmetric to the one above. Current parameter is $\theta_T = (\beta_T, \sigma_T)$ so propose $u \sim J_{T,S}(u|\theta_T) = P_S(u)$ and use function

$$g_{T,S}(\theta_T, u) = (u_1, \exp\{u_2\}, \beta_T, \sigma_T) = (\beta_S^*, \sigma_S^*, u_1^*, \exp\{u_2^*\})$$

to match the dimension. The acceptance probability is then

$$\alpha_{S,T} = \frac{L(\boldsymbol{\rho}; \theta_S^*, M_S) p(\theta_S^* \mid M_S) J_{S,T}(u^*)}{L(\boldsymbol{\rho}; \theta_T, M_T) p(\theta_T \mid M_T) J_{T,S}(u)} \exp\{u_2\}.$$

Moving from M_S to M_{ST}. Current parameter is $\theta_S = (\beta_S, \sigma_S)$ so need an auxiliary

variable u relating to the other 5 parameters of M_{ST}. This can be proposed from $J_{S,ST}(u|\theta_S) = J_{S,ST}(u)$, a 5-dimensional Gaussian derived from the 7-dimensional Gaussian $P_{ST}(u)$, where u is the vector resulting from taking away the first and third elements of (17). Then, use the function

$$g_{S,ST}(\theta_S, u) = \left(\beta_S, u_1, \sigma_S, q_\sigma^{-1}(u_2), q_\phi^{-1}(u_3), q_\phi^{-1}(u_4), q_\pi^{-1}(u_5)\right) = \theta_{ST}^*$$

to do the dimension matching. In other words, (β_S, σ_S) remain the same while the five remaining parameters of M_{ST} are being generated. It is straightforward to show that the necessary Jacobian determinant is given by

$$\left|\frac{\partial g_{S,ST}(\theta_S, u)}{\partial\left(\theta_S, u\right)}\right| = e^{u_2} n^2 \prod_{i=3}^{5} \frac{1}{2 + e^{u_i} + e^{-u_i}} = G_{S,ST} \tag{18}$$

so that the acceptance probability is:

$$\alpha_{S,ST} = \frac{L(\boldsymbol{\rho}; \theta_{ST}^*, M_{ST}) p(\theta_{ST}^* \mid M_{ST})}{L(\boldsymbol{\rho}; \theta_S, M_S) p(\theta_S \mid M_S) J_{S,ST}(u)} \times G_{S,ST}.$$

Note that there is no auxiliary variable needed for the opposite move from M_{ST} to M_S as described below.

Moving from M_{ST} to M_S. Current parameter is θ_{ST} so just take its first and third elements to get $\theta_S^* = (\beta_S^*, \sigma_S^*)$. The acceptance probability is

$$\alpha_{ST,S} = \frac{L(\boldsymbol{\rho}; \theta_S^*, M_S) p(\theta_S^* \mid M_S) J_{S,ST}(u^*)}{L(\boldsymbol{\rho}; \theta_{ST}, M_{ST}) p(\theta_{ST} \mid M_{ST})}.$$

where u^* is obtained from current θ_{ST} as

$$u^* = \left(\beta_T, q_\sigma(\sigma_T), q_\phi(\phi_S), q_\phi(\phi_T), q_\pi(\pi_S)\right).$$

Moving from M_T to M_{ST}. This move is symmetric to the move from M_S to M_{ST}. Current parameter is $\theta_T = (\beta_T, \sigma_T)$ so need u relating to the other 5 parameters of M_{ST}. This can be proposed from $J_{T,ST}(u)$, the 5-dimensional Gaussian derived from $P_{ST}(u)$, where u is the vector resulting from taking away the second and fourth elements of (17). Then, use the function

$$g_{T,ST}(\theta_T, u) = \left(u_1, \beta_T, q_\sigma^{-1}(u_2), \sigma_T, q_\phi^{-1}(u_3), q_\phi^{-1}(u_4), q_\pi^{-1}(u_5)\right) = \theta_{ST}^*$$

whose Jacobian determinant $G_{T,ST}$ is the same as (18). The acceptance probability is:

$$\alpha_{T,ST} = \frac{L(\boldsymbol{\rho}; \theta_{ST}^*, M_{ST}) p(\theta_{ST}^* \mid M_{ST})}{L(\boldsymbol{\rho}; \theta_T, M_T) p(\theta_T \mid M_T) J_{T,ST}(u)} \times G_{T,ST}.$$

Moving from M_{ST} to M_T. Current parameter is θ_{ST} so just take its second and

fourth elements to get $\theta_T^* = (\beta_T^*, \sigma_T^*)$. The acceptance probability is

$$\alpha_{ST,T} = \frac{L(\boldsymbol{\rho}; \theta_T^*, M_T) p(\theta_T^* \mid M_T) J_{T,ST}(u^*)}{L(\boldsymbol{\rho}; \theta_{ST}, M_{ST}) p(\theta_{ST} \mid M_{ST})}.$$

where u^* is obtained from current θ_{ST} as

$$u^* = (\beta_S, q_\sigma(\sigma_S), q_\phi(\phi_S), q_\phi(\phi_T), q_\pi(\pi_S)) .$$

Within-model moves. All internal moves are done using MH, with proposal distributions $P_S(\cdot)$, $P_T(\cdot)$, $P_{ST}(\cdot)$ obtained from individual MCMC runs for each of the three models.

4 Model application

4.1 EN4 Reanalysis data

The available observations are in the form of a reanalysis product (EN4, Good et al., 2013). EN4 ingests quality controlled ocean observations of temperature and salinity for each month and merges them onto a regular grid. It then infills the remainder by optimal interpolation using fixed horizontal and vertical decorrelation length scales and relaxing to its own climatology with an e-folding timescale of 9.5 months. The resulting time series are values of ocean density $\rho(t)$ for $t = 1, \ldots, 115$ years, corresponding to the time period 1900-2014 inclusive.

Three MCMC chains were run (in parallel)—for each of the models M_S, M_T and M_{ST}—for 100K iterations. The burn-in for M_S and M_T was 10K while for M_{ST} it was 50K to ensure convergence. This was assessed by visual plots and the Gelman and Rubin $\hat{R}$ multi-chain diagnostic (Gelman et al., 2013). Subsequently, the reversible jump algorithm described earlier was run for 50K iterations (once for each of the three chains), using the adapted proposal distributions from the individual model runs. Figure 3 shows a trace plot of the logarithm of the posterior distribution, computed at each MCMC sample. The log-posterior is a summary of all model parameters and the plot indicates overall convergence (individual plots also assessed but not shown).

The posterior probability for model M_{ST} was $p(M_{ST}|\boldsymbol{\rho}) = 0.9999$ indicating very strong evidence of a regime switching scenario for this data set (only 15 out of 150K samples indicated model M_S, while none for M_T). As such, we consider model M_{ST} as the most appropriate for this data set, and proceed to ensure an adequate fit to the data.

4.1.1 Comparison with HMM

As mentioned in Section 3, the HSMM is an extension of the HMM, to allow for holding time distributions other than Geometric (at the expense of computing time).

It is therefore of interest to compare the way that the fitted HSMM (with zero-truncated Poisson distributions) is different from the HMM. To that end, an HMM was implemented to the EN4 data. In-sample model checking (see subsequent section), indicated that the two models captured the data in a similar manner. However, we also conducted out of sample prediction, to test the temporal structure in each model. This was done by removing m years from the end of the time series and then predicting the $\rho(t)$ for that time period. This was performed for $m = 5, \ldots, 60$ years, noting that the total time period was 115 years.

Figure 2 shows the root mean squared (RMSE) and the mean absolute error (MAE) plotted against m. Both plots indicate little difference between the HSMM and the HMM (although HSMM marginally better), for up to about 50 years out of sample prediction. Beyond this, the HMM exhibits a fairly sharp decline in predictive power, particularly in the RMSE. This is likely due to the nature of the Geometric holding time distribution where the variance is proportional to the square of the mean, unlike the truncated Poisson in the HSMM where the mean equals the variance. For the same mean holding time, the HMM is likely to predict more extreme holding times. Given the application to ocean density where we a-priori believe that the state switching mechanism is present (i.e. holding time periods of extreme length are less likely), we find the HSMM a more appropriate modelling assumption.

4.1.2 Model checking

Model checking is performed by comparing the data $\boldsymbol{\rho}$ with associated predictions of $\boldsymbol{\rho}$ from the model. The predictive distribution for each data point $\rho(t)$ is

$$p(\rho(t)|\boldsymbol{\rho}) = \int_{\boldsymbol{\theta}} \sum_{C_T=S}^{T} p(\rho(t) \mid C(t), \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\rho}) p(C(t)|\boldsymbol{\rho}) d\boldsymbol{\theta}. \tag{19}$$

To obtain samples from (19), we must first run the backward algorithm outlined in Section 3.3.1 to obtain the predictive probability $p(C(t) = S|\boldsymbol{\rho})$ of the latent chain being in state S at time t. We can then sample state S (and thus T) as a Bernoulli realisation. Furthermore, we obtain a sample of $\boldsymbol{\theta}$ (the parameters of M_{ST}) by taking one of the MCMC samples from the posterior $p(\boldsymbol{\theta}|y)$. Finally using the sample of the state $C(t)$ and the sample of $\boldsymbol{\theta}$, we simulate a value from the appropriate conditional model (9). Repeating this process for each MCMC sample i provides simulated values from (19), and thus we can construct predicted data sets $\boldsymbol{\rho}^{(i)}$.

As such, we can perform posterior predictive model checking (Gelman et al., 2013) by enquiring as to whether any of the observed $\rho(t)$ are an extreme with respect to (19). To succinctly summarise such discrepancies, one can look at summary statistics of the data and compare them with their respective posterior distributions computed from predictive samples. Figure 4 shows the predictive distribution of the sample mean and standard deviation, as well as the lower and upper sample quartiles. The observed quantities of those are given in vertical lines, indicating that none is an extreme values alluding to an adequate model fit. There is some evidence that the

lower quartile in slightly overestimated while the upper one is underestimated, but the associated tail area probabilities (0.22 and 0.10 respectively) are not too extreme.

We can also look at how well the marginal distribution of $\rho(t)$ was captured as well as individual points. The left panel in Figure 5 shows posterior predictive means, of predictions sorted in ascending order, plotted against the sorted observations, along with 99% prediction intervals. This can be interpreted like a Q-Q plot, indicating that the marginal distribution of $\rho(t)$ is well captured, with the exception of the maximum data point which is just outside the upper endpoint of the interval. The right panel in Figure 5, shows posterior predictive means for each data point, plotted against the observed $\rho(t)$, along with corresponding 99% prediction intervals. The plot indicates the points are well captured, with the exception of the two larger values. Summarising, the model describes the marginal distribution of the data and individual points well, with slight underestimation of the very extremes. Given the goal here is less about describing extremes but rather to capture the regime modulated relationship of $\rho(t)$ with $\rho_S(t)$ and $\rho_T(t)$, we consider this discrepancy acceptable.

A more succinct way to compare observations with model predictions, is to consider a measure that involves both data and parameters. A proposed option (Economou et al., 2014) is the deviance $D(\boldsymbol{\theta}, \boldsymbol{\rho})$, defined as minus twice the logarithm of the likelihood (15). For each MCMC sample i from $p(\boldsymbol{\theta}|\boldsymbol{\rho})$, the deviance $D(\boldsymbol{\theta}^{(i)}, \boldsymbol{\rho})$ for the observed data can be computed, as can the deviance for the predictions $D(\boldsymbol{\theta}^{(i)}, \boldsymbol{\rho}^{(i)})$. If the posterior of the "observed" deviance is an extreme with respect to the "predicted" one, then this indicates that data generated from the model are very different to the observed data. Figure 6 shows trace plots of both $D(\boldsymbol{\theta}^{(i)}, \boldsymbol{\rho})$ and $D(\boldsymbol{\theta}^{(i)}, \boldsymbol{\rho}^{(i)})$ indicating the observed data set $\boldsymbol{\rho}$ is not an extreme with respect to predictions. The probability $\Pr(D(\boldsymbol{\theta}, \tilde{\boldsymbol{\rho}}) > D(\boldsymbol{\theta}, \boldsymbol{\rho}))$—where $\tilde{\boldsymbol{\rho}}$ denotes predictions—is estimated to be 0.64 by the analogous proportion of samples used to produce Figure 6.

Finally, Figure 7 shows the sample autocorrelation function of the data $\boldsymbol{\rho}$ for lags of up to 15 years. Added to the plot are the posterior means and 95% credible intervals of the autocorrelation at each lag, computed from the predictive samples. All the sample values are within the credible intervals indicating that the model adequately captures the autocorrelation in the data, justifying the choice of holding time distribution (which ultimately controls the autocorrelation in the marginal distribution of an HSMM).

4.1.3 Interpretation

Table 2 shows the posterior means for $\boldsymbol{\theta}$ as well as 95% credible intervals. The magnitude of β_S implies that the salinity regime is more dominant than the temperature regime. In terms of regime persistence, the mean holding time for the salinity regime is 11.2 years while for the temperature regime it is just 1.6 years. These timescales are consistent with the prevailing view that deep convection in the Labrador Sea is episodic, sometimes not occurring for a decade at a time (Yashayaev and Loder, 2016), but that when it occurs it is controlled by temperature rather than salinity. Figure 8 shows a plot of $P(C(t) = S \mid \boldsymbol{\rho})$, the posterior probability of being in regime

S, plotted against t. The most likely regime sequence (obtained using the method described in Section 3.3.1) is also shown in Figure 8. Note that in this case the most likely regime sequence coincides with one that would be obtained by just using the marginal distribution $P(C(t) = S \mid \boldsymbol{\rho})$, however in general this will not be the case. The plot indicates that the temperature state only holds for very short periods of 1-2 years. To further visualise this, Figure 9 shows plots of $\rho(t)$ against $\rho_S(t)$ ($\rho_T(t)$) for the subset of data classified as most likely being in regime S (T). Interestingly, the values of $\rho(t)$ classified as being driven by regime T are all effectively zero, implying that the temperature regime is one where ocean density varies very little about its mean (noting also that $\rho_T(t)$ is much less variable than $\rho_S(t)$). It is possible that this is linked to the stronger damping of annual temperature variability (at the atmosphere-ocean boundary), effectively equalizing temperature anomalies. In addition, the much longer holding time for the (more weakly damped) salinity regime may provide more time for a stronger relationship between $\rho_S(t)$ and $\rho(t)$. Although outside the scope of this study, future work should investigate the structure of the salinity mean holding time and its physical evolution.

In summary, there is compelling evidence to suggest that for the particular data set, ocean density variability in the Labrador Sea is governed by salinity/temperature regime changes. The salinity regime is more prevalent in terms of holding time but also in terms of magnitude of the driving signal. This is consistent with recent work that investigated a reanalysis system that was strongly constrained by EN4 data, though a competing reanalysis system did not show this behaviour (Menary and Hermanson, 2018). Although individual deep convective events in the Labrador Sea are likely temperature controlled (Yashayaev and Loder, 2017), there is clearly scope for changes in the background stratification to be modulated by salinity, as shown in reanalyses and climate models (Menary et al., 2015). Unfortunately, a lack of subsurface observations, particularly of salinity, preclude complete knowledge of recent real world decadal variability. As such, we apply our methodology to a large suite of climate models, in order to better understand this uncertainty.

4.2 Application to 43 climate models

As mentioned in the introduction, one of the goals here is to use the model to investigate whether various coupled general circulation models (CGCMs) of the climate, exhibit the same regime-changing behaviour observed in the EN4 data. CGCMs can be run for much longer time intervals, and so should be able to provide a better understanding of the decadal behaviour in ocean density variability.

The model described in Section 4.1 was applied to 43 pre-industrial control simulations from CGCMs that participated in the fifth coupled model intercomparison project (CMIP5, Taylor et al., 2012). These simulations aim to recreate an equilibrium climate (prior to the secular trend that is now evident) using interannualy invariant external forcings (e.g. solar, greenhouse gas, etc) appropriate for pre-industrial times. Each control simulation was at least 200 years in length. They represent different approaches to simulating this pre-industrial climate and by comparing them it

is possible to investigate the strength of internal variability in the climate system.

Adequate model fit was ensured by looking at tail area probabilities derived from the model checking measures described in Figures 4 and 6. Tables 3, 4 and 5 in the Appendix, summarise the results in terms of the most likely model (M_S, M_T, M_{ST}) for each CGCM (second column), as chosen by reversible jump. The other columns show the posterior mean and 95% credible interval for the parameters of the most likely model. Quite evidently there is much variability across the various CGCMs. Starting however from common features, we see that only 8 out of 43 CGCMs exhibit a non-regime changing scenario. Out of those 8, just one is solely temperature driven (i.e. model M_T), while the other 7 are all salinity driven (model M_S). This preference for salinity, over temperature, perhaps reflects the relatively low ocean resolution of these climate models, which is of order 1° longitude/latitude, and their associated mean state biases (Menary et al., 2015). Simulations with higher resolution models as part of the ongoing sixth coupled model intercomparison project (CMIP6) may reveal different behaviour.

Most CGCMs are able to simulate a regime changing ocean density in line with the reanalysis data EN4, albeit with varying degrees of temporal persistence of the regimes. Overall, the salinity state has larger mean holding times across CGCMs, much like EN4. Some CGCMs however have a temperature regime that lasts longer on average than the salinity one.

4.2.1 Relationship with the Atlantic Meridional Overturning Circulation (AMOC)

The statistical model parameters, in particular the ones relating to the holding times of each regime (ϕ_S and ϕ_T), provide a way of quantifying the temporal regime behaviour of each CGCM. As noted previously, density in the Labrador Sea is important for the strength and variability in the climatically relevant AMOC. Thus, an important question is whether the strength of the AMOC in CGCMs is systematically related to the preference for one density regime or the other. To investigate this question, we propose a single metric that is a measure of the relative regime persistence, defined as $\log(\phi_T/\phi_S)$, where larger (smaller) values suggest a more temperature (salinity) dominated density regime.

From Figure 10, it can be seen that the AMOC strength in complex CGCMs is indeed linked to their preference for one density regime over another. In this analysis, we are limited to the intersection of those climate modelling centres that uploaded AMOC streamfunction data to the CMIP5 archive (see Table 1 of Menary and Wood (2017)) and those CGCMs for which both regimes are active (i.e. CGCMs for which M_{ST} was the most likely model), which results in a reduced subset of 15 CGCMs. We define the AMOC at two latitudes: 26.5°N, which is the latitude of the recently deployed 'RAPID-MOCHA' array (Cunningham et al., 2007), and 45°N, which represents the boundary of the subtropical and subpolar gyres in the North Atlantic. At both latitudes, a strong linear relationship can be seen. CGCMs that have increasingly temperature-driven density variability in the Labrador Sea tend to have a stronger

AMOC in the mean, with correlations of 0.79 at 26.5°N and 0.66 at 45°N (after removing one outlier). The outlier at the 45°N AMOC, is the GISS-E2-R model. This model was also found to be an outlier in recent climate modelling studies of the North Atlantic subpolar gyre (Menary and Wood, 2017; Sgubin et al., 2017). Further investigation and analysis of the results, as well as the physical implications will be the focus of future work.

In previous work, a similar index of the density driver in CGCMs was constructed, denoted $\rho_{TorScontrol}$, for which larger values implied an increasing dominance of temperature variability (Menary et al., 2015). This index was simply the difference between two regression coefficients (density versus density due to temperature and density due to salinity, respectively) and took no account of the switching behaviour allowed for here, nor the relative holding times in one or the other regime. Nonetheless, for the 36 models that show regime change (and thus $\log(\phi_T/\phi_S)$ can be defined) we can compare the two methods. We find that the cross-model correlation between $\rho_{TorScontrol}$ and $\log(\phi_T/\phi_S)$ is 0.83, which provides confidence that the two, independent approaches are measuring a similar phenomenon. However, unlike the previous method, the model we present here provides much further scope for understanding the nature of density variability in these climate models—and thus the real world ocean—in this particularly important region.

All the data and associated `R` (R Core Team, 2017) code used to implement the models in this section, are provided in the supplementary material.

5 Discussion

A modelling framework for inferring regime changes in ocean density was presented. This was based on the concept of hidden semi-Markov models, which is a natural framework for describing discrete but unobserved state changes in a system. The model was implemented in the Bayesian framework and as such, reversible jump MCMC was employed to choose between three candidate models for describing ocean variability. RJMCMC has only been applied in the context of HMMs so far, and in the special case that the parameters of conditional model can be interpreted across models with varying number of latent states. Here we present a formulation that is not constrained to this case, and is thus potentially more widely applicable.

The model was applied to a reanalysis (a proxy for observed) data set of ocean density in the Labrador Sea. Model checking based on the predictive distribution of the data was performed to ensure model adequacy. The results suggested a regime changing setting with a long lasting salinity and a shorter lasting temperature regime. Further applying the model to 43 free running CGCMs indicated that most of these evince a similar regime changing scenario as the reanalysis data. Some CGCMs however manifested a single regime case.

The conditional models utilised here were specific to the application of ocean density variability—in fact they were derived from physical arguments. Nevertheless, the

framework presented here is potentially applicable to any HSMM with any conditional model for each latent state. In particular, the various conditional models across models with varying number of states do not need to be in any way nested—and in the case of HSMMs this avoids self transitions (and thus non-identifiability in the holding times). Therefore, the method presented here can be used as simply a way for choosing the HSMM with the optimal number of states.

The requirement that each of the possible models is ran first may be seen as inelegant and one that increases computational strain. On the other hand, this maximises flexibility and does not pose any constraints in the definition of the conditional models. In our experience, the final reversible jump run after all models have been fitted is relatively very cheap since the proposal distributions are established from the single runs.

References

Ba, J., Keenlyside, N. S., Latif, M., Park, W., Ding, H., Lohmann, K., Mignot, J., Menary, M., Otterå, O. H., Wouters, B., et al. (2014). A multi-model comparison of Atlantic multidecadal variability. Climate Dynamics, 43(9-10):2333–2348.

Bellone, E., Hughes, J. P., and Guttorp, P. (2000). A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. Climate Research, 15:1–12.

Collins, M. (2002). Climate predictability on interannual to decadal time scales: The initial value problem. Climate Dynamics, 19(8):671–692.

Cunningham, S. A., Kanzow, T., Rayner, D., Baringer, M. O., Johns, W. E., Marotzke, J., Longworth, H. R., Grant, E. M., Hirschi, J. J. M., Beal, L. M., Meinen, C. S., and Bryden, H. L. (2007). Temporal variability of the Atlantic meridional overturning circulation at 26.5°N. Science, 317(5840):935–938.

Delworth, T., Manabe, S., and Stouffer, R. J. (1993). Interdecadal variations of the thermohaline circulation in a coupled ocean-atmosphere model. Journal of Climate, 6(11):1993–2011.

Economou, T., Bailey, T. C., and Kapelan, Z. (2014). Mcmc implementation for bayesian hidden semi-markov models with illustrative applications. Statistics and Computing, 24(5):739–752.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). Bayesian data analysis. Chapman and Hall/CRC.

Good, S. A., Martin, M. J., and Rayner, N. A. (2013). En4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. Journal of Geophysical Research: Oceans, 118(12):6704–6716.

Green, P. J. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika, 82:711–732.

Guedon, Y. (2003). Estimating hidden semi-Markov chains from discrete sequences. Journal of Computational and Graphical Statistics, 12(3):604–639.

Hu, C. and Gruttola, V. D. (2007). Joint modeling of progression of hiv resistance mutations measured with uncertainty and failure time data. Biometrics, 63(1):60–68.

Hughes, J. P., Guttorp, P., and Charles, S. P. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. Journal of the Royal Statistical Society: Series C (Applied Statistics), 48(1):15–30.

Jouyaux, C., Richardson, S., and Longini, I. (2000). Modeling markers of disease progression by a hidden Markov process: Application to characterizing cd4 cell decline. Biometrics, 56(3):733–741.

Kozumi, H. (2000). Bayesian analysis of discrete survival data with a hidden Markov chain. Biometrics, 56(4):1002–1006.

Menary, M. B. and Hermanson, L. (2018). Limits on determining the skill of north atlantic ocean decadal predictions. Nature Communications, 9(1694).

Menary, M. B., Hermanson, L., and Dunstone, N. J. (2016). The impact of labrador sea temperature and salinity variability on density and the subpolar amoc in a decadal prediction system. Geophysical Research Letters, 43(23):12,217–12,227.

Menary, M. B., Hodson, D. L. R., Robson, J. I., Sutton, R. T., Wood, R. A., and Hunt, J. A. (2015). Exploring the impact of cmip5 model biases on the simulation of north atlantic decadal variability. Geophysical Research Letters, 42(14):5926–5934.

Menary, M. B. and Wood, R. A. (2017). An anatomy of the projected North Atlantic warming hole in CMIP5 models. Climate Dynamics, pages 1–18.

R Core Team (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Robert, C. P., Rydèn, T., and Titterington, D. M. (2000). Bayesian inference in hidden markov models through the reversible jump markov chain monte carlo method. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62(1):57–75.

Rydén, T., Terasvirta, T., and Asbrink, S. (1998). Stylized facts of daily return series and the hidden Markov model. Journal of Applied Econometrics, 13:217–244.

Sansom, J. and Thomson, P. (2001). Fitting hidden semi-Markov models to breakpoint rainfall data. Journal of Applied Probability, 38A:142–157.

Scheidegger, A. (2018). adaptMCMC: Implementation of a Generic Adaptive Monte Carlo Markov Chain Sampler. R package version 1.3.

Scott, S. (2002). Bayesian methods for hidden Markov models: recursive computing in the 21st century. Journal of the American Statistical Association, 97:337–351.

Sgubin, G., Swingedouw, D., Drijfhout, S., Mary, Y., and Bennabi, A. (2017). Abrupt cooling over the north atlantic in modern climate models. Nature Communications, 8.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2012). An overview of cmip5 and the experiment design. Bulletin of the American Meteorological Society, 93(4):485–498.

Tokdar, S., Xi, P., Kelly, R., and Kass, R. (2010). Detection of bursts in extracellular spike trains using hidden semi-Markov point process models. Journal of Computational Neuroscience, 29:203–212.

Vihola, M. (2012). Robust adaptive metropolis algorithm with coerced acceptance rate. Statistics and Computing, 22(5):997–1008.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory, 13:260–269.

Yashayaev, I. and Loder, J. W. (2016). Recurrent replenishment of labrador sea water and associated decadal-scale variability. Journal of Geophysical Research: Oceans, 121(11):8095–8114.

Yashayaev, I. and Loder, J. W. (2017). Further intensification of deep convection in the labrador sea in 2016. Geophysical Research Letters, 44(3):1429–1438.

Yu, S.-Z. (2010). Hidden semi-Markov models. Artificial Intelligence, 174:215–243.

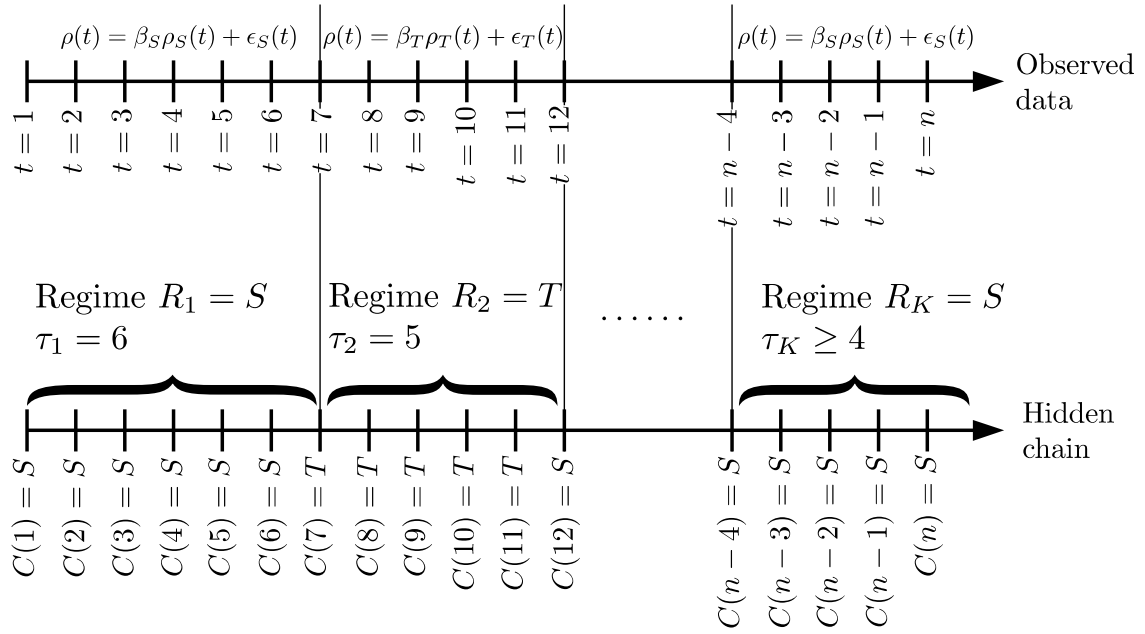

Figure 1: Schematic showing a particular realisation of the hidden semi-Markov model for ocean density given by (9)–(10).

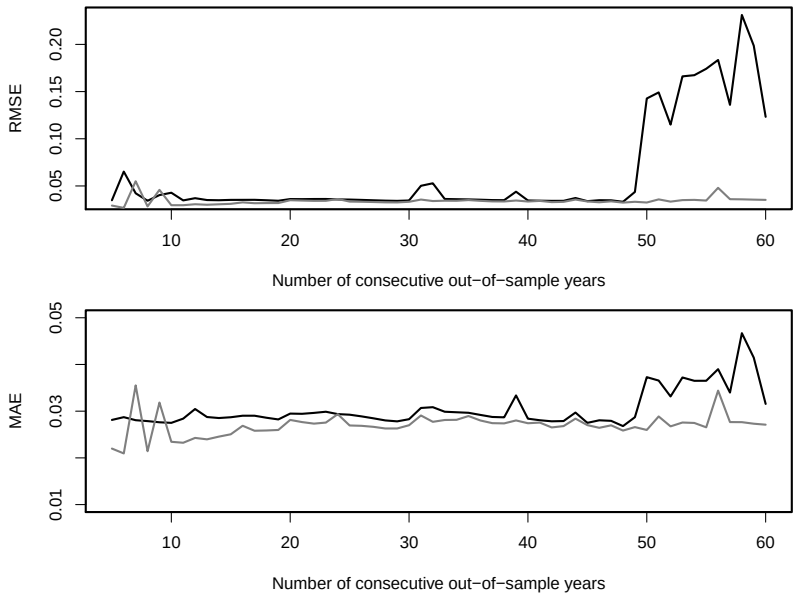

Figure 2: Out of sample predictions (root mean squared error on top and man absolute error on bottom) of the EN4 data, comparing the HSMM (grey lines) with the HMM (black lines). The x-axis represent the number of consecutive years used for prediction at the end of the observation window.

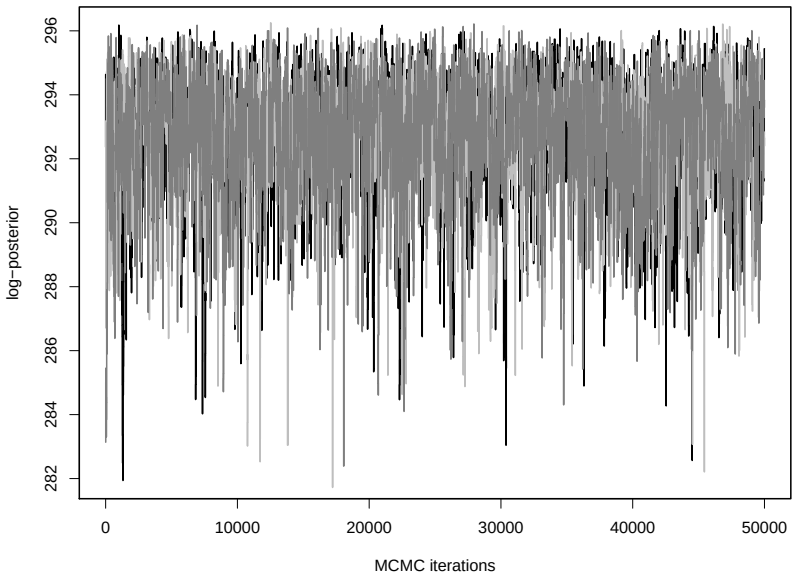

Figure 3: Trace plot of the logarithm of the posterior distributed evaluation at each sample from the joint posterior.

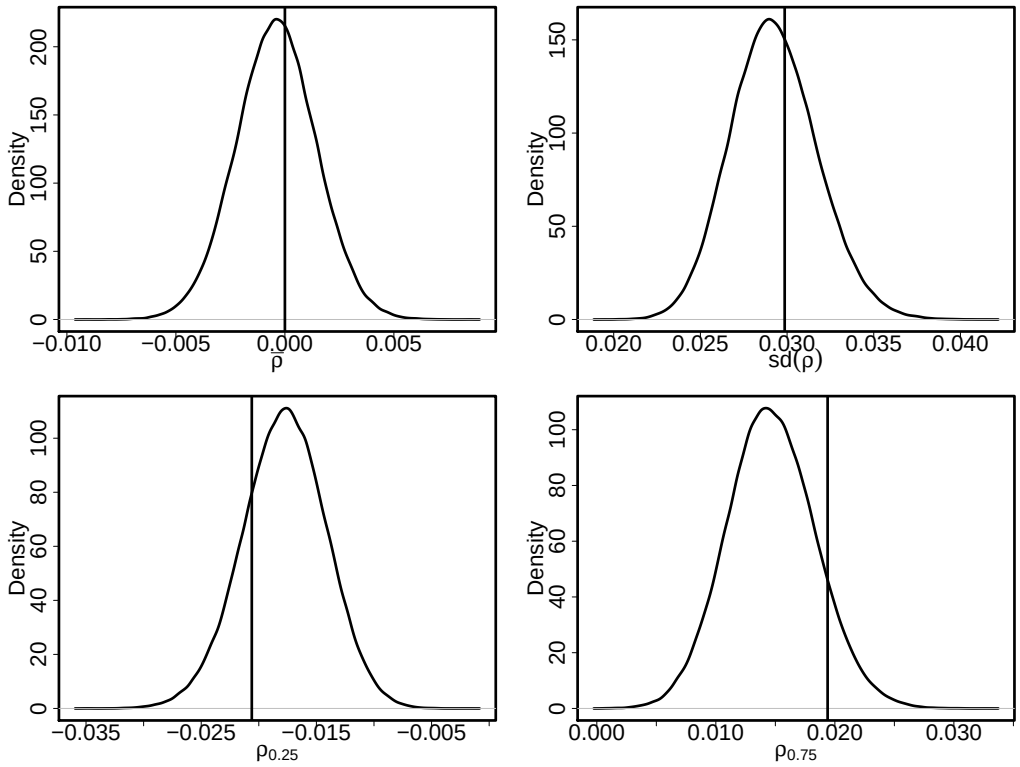

Figure 4: Top left: posterior (predictive) distribution of the sample mean $\bar{\rho}$. Top right: posterior of the sample standard deviation. Bottom left/right: posterior of the lower/upper quartile. The equivalent sample quantities are shown by vertical lines.

Appendix

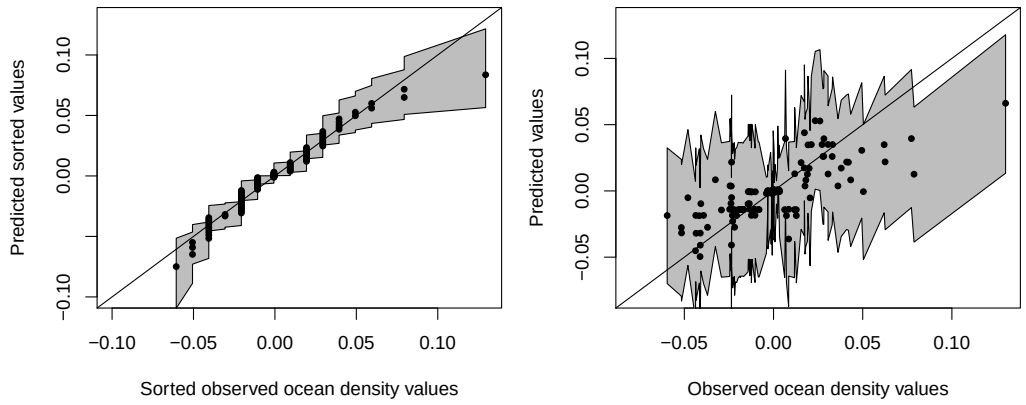

Figure 5: Left: Predicted values sorted in ascending order plotted against the sorted observed values. Right: Predicted values (means of posterior predictive distributions) plotted against observed. Both plots include 99% prediction intervals.

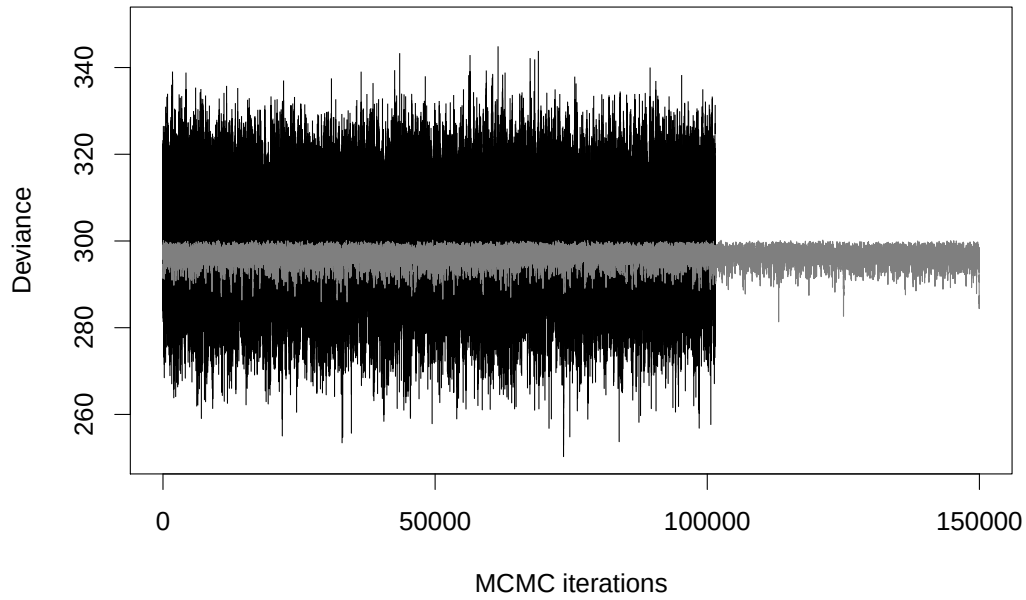

Figure 6: Trace plot of the "observed" deviance $D(\boldsymbol{\theta}^{(i)}, \boldsymbol{\rho})$ against i in grey; and of the "predicted" $D(\boldsymbol{\theta}^{(i)}, \boldsymbol{\rho}^{(i)})$ in black.

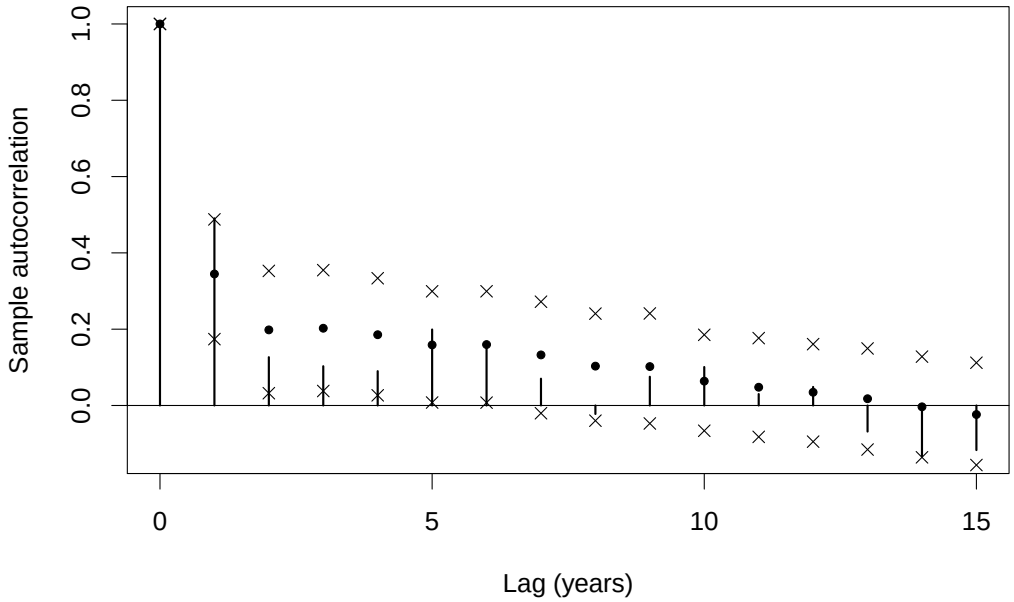

Figure 7: Sample autocorrelation function of the data (vertical lines), along with the posterior mean of the autocorrelation at each lag(points) and associated 95% credible intervals (crosses).

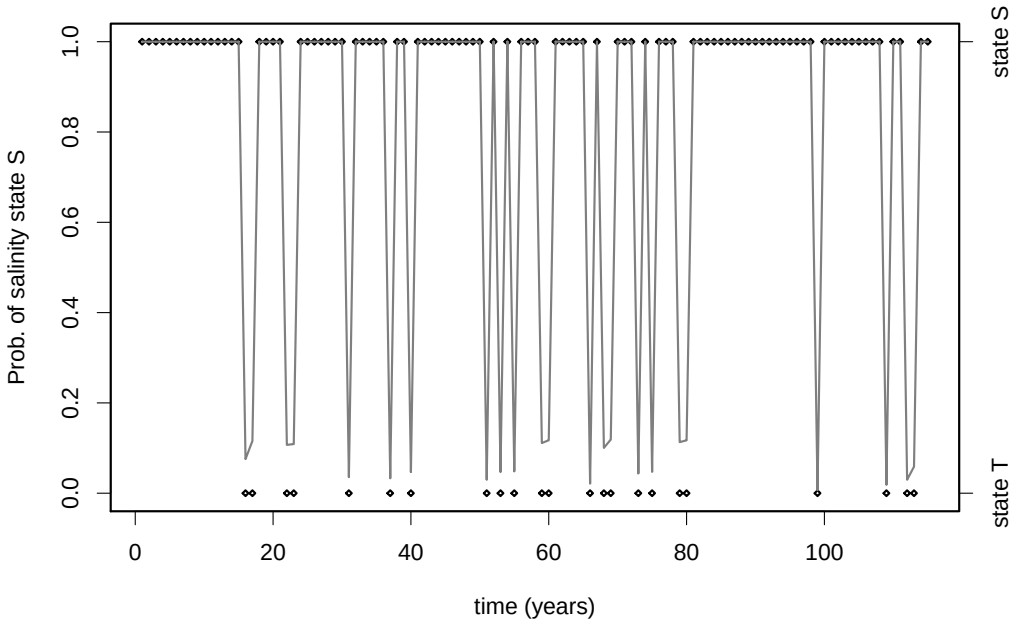

Figure 8: Plot of $P(C(t) = S \mid \boldsymbol{\rho})$ against t (years). The most likely state sequence is also added using a right hand y-axis and diamond symbols

.

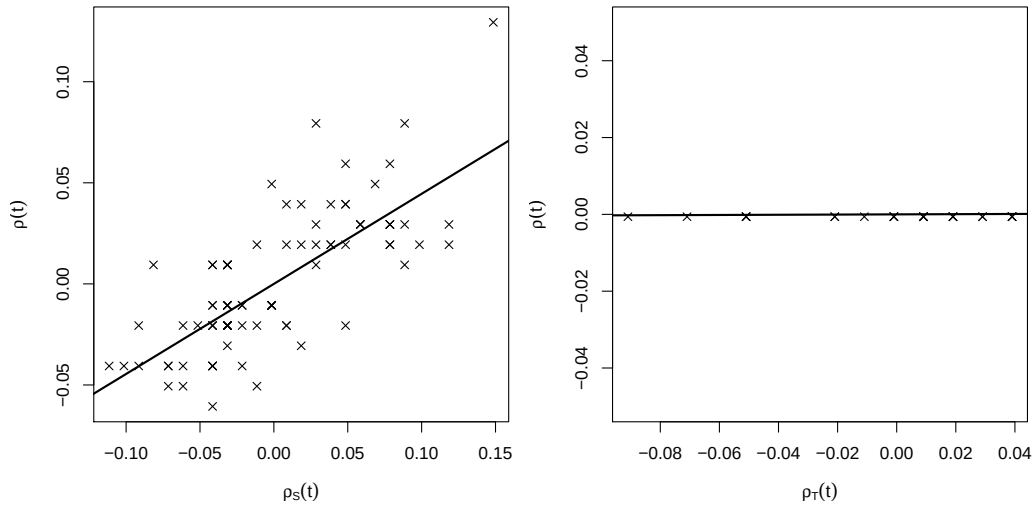

Figure 9: Left: plot of $\rho(t)$ against $\rho_S(t)$ for the subset of data where the most likely state is S, and the estimated relationship $\beta_S \rho_S(t)$. Right: corresponding plot for regime T.

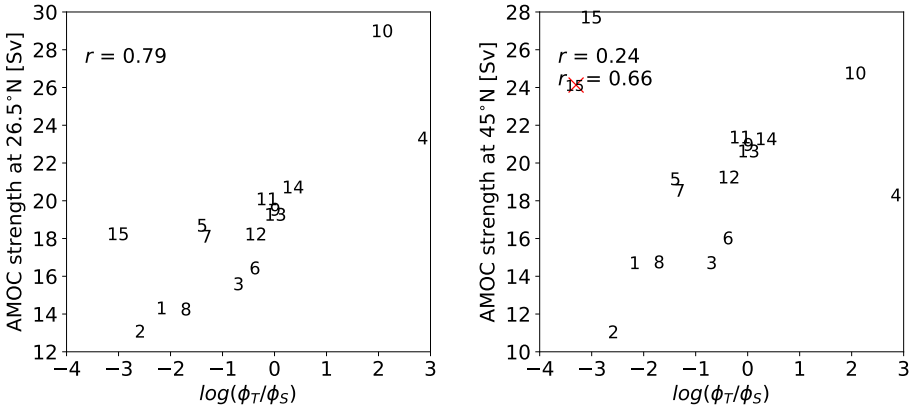

Figure 10: Relationship between relative regime persistence and the mean AMOC strength at 26.5°N (left) and 45°N (right) in CGCMs. AMOC strength is measured in Sverdrups [1 Sv = $10^6 m^3/s$].

Table 1: CGCM legend

Number	CGCM
1	CanESM2
2	CNRM-CM5
3	CNRM-CM5-2
4	FGOALS-g2
5	MPI-ESM-LR
6	MPI-ESM-MR
7	MPI-ESM-P
8	MRI-CGCM3
9	CCSM4
10	NorESM1-ME
11	CESM1-BGC
12	CESM1-CAM5
13	CESM1-FASTCHEM
14	CESM1-WACCM
15	GISS-E2-R

Table 2: Parameter estimates for model M_{ST} fitted to the EN4 reanalysis data set.

Parameter	Posterior Mean	95% Cr.I.
β_S	0.446	$[0.367, 0.527]$
β_T	0.003	$[-0.008, 0.013]$
σ_S	0.022	$[0.019, 0.025]$
σ_T	0.0007	$[0.0004, 0.001]$
ϕ_S	11.205	$[7.537, 16.845]$
ϕ_T	0.970	$[0.257, 1.957]$
π_S	0.665	$[0.158, 0.986]$

CGCM	Most likely model	β_S	σ_S	β_T	σ_T	ϕ_S	ϕ_T
bcc.csm1.1	M_S $(p = 0.96)$	0.54	0.014	-	-	-	-
bcc.csm1.1	M_{ST} $(p = 0.04)$	[0.51,0.57]	[0.013,0.015]	-	-	-	-
BNU.ESM	M_{ST} $(p = 1)$	0.93	0.019	0.97	0.017	23.07	3.82
BNU.ESM		[0.88,0.99]	[0.018,0.020]	[0.75,1.16]	[0.012,0.022]	[19.69,26.12]	[2.39,5.64]
CMCC.CESM	M_{ST} $(p = 1)$	0.73	0.012	1.30	0.020	94.65	6.57
CMCC.CESM		[0.70,0.76]	[0.011,0.014]	[0.78,1.76]	[0.013,0.035]	[79.90,110.70]	[3.13,11.65]
CMCC.CMS	M_S $(p = 0.99)$	0.73	0.015	-	-	-	-
CMCC.CMS	M_{ST} $(p = 0.01)$	[0.71,0.75]	[0.014,0.016]	-	-	-	-
CNRM.CM5.2	M_{ST} $(p = 1)$	0.60	0.025	0.80	0.026	8.27	4.11
CNRM.CM5.2		[0.53,0.67]	[0.023,0.029]	[0.65,0.94]	[0.022,0.030]	[5.97,13.16]	[2.87,5.77]
ACCESS1.3	M_{ST} $(p = 1)$	0.38	0.035	0.63	0.036	15.68	3.41
ACCESS1.3		[0.35,0.42]	[0.032,0.039]	[0.48,0.78]	[0.023,0.051]	[11.08,21.50]	[2.31,4.89]
FIO.ESM	M_{ST} $(p = 1)$	0.63	0.017	0.60	0.019	6.03	4.05
FIO.ESM		[0.58,0.69]	[0.015,0.019]	[0.52,0.69]	[0.017,0.022]	[4.47,9.23]	[3.10,5.64]
inmcm4	M_S $(p = 1)$	0.66	0.042	-	-	-	-
inmcm4		[0.65,0.68]	[0.040,0.045]	-	-	-	-
IPSL.CM5A.MR	M_S $(p = 0.93)$	0.78	0.016	-	-	-	-
IPSL.CM5A.MR	M_{ST} $(p = 0.07)$	[0.76,0.81]	[0.015,0.017]	-	-	-	-
FGOALS.g2	M_{ST} $(p = 0.99)$	1.15	0.028	0.91	0.014	5.59	96.12
FGOALS.g2	M_T $(p = 0.01)$	[0.78,1.46]	[0.021,0.037]	[0.89,0.94]	[0.013,0.015]	[3.58,7.95]	[69.79,108.16]
HadGEM2.CC	M_{ST} $(p = 1)$	0.51	0.026	0.96	0.036	10.66	5.41
HadGEM2.CC		[0.42,0.60]	[0.019,0.032]	[0.60,1.22]	[0.026,0.049]	[6.06,17.03]	[3.45,9.56]
MPI.ESM.LR	M_{ST} $(p = 1)$	0.67	0.021	0.73	0.022	14.37	3.55
MPI.ESM.LR		[0.63,0.71]	[0.019,0.022]	[0.62,0.85]	[0.019,0.026]	[9.32,17.36]	[2.70,4.51]
MPI.ESM.P	M_{ST} $(p = 1)$	0.66	0.023	0.55	0.022	15.48	4.16
MPI.ESM.P		[0.66,0.73]	[0.022,0.024]	[0.55,0.77]	[0.018,0.025]	[13.33,18.16]	[3.20,5.23]
MRI.ESM1	M_{ST} $(p = 0.99)$	0.64	0.022	0.61	0.042	30.33	5.22
MRI.ESM1	M_S $(p = 0.01)$	[0.56,0.72]	[0.019,0.025]	[0.25,1.11]	[0.028,0.063]	[25.00,37.81]	[2.61,8.54]
GISS.E2.H.CC	M_{ST} $(p = 1)$	0.54	0.017	-0.03	0.0126	16.25	4.40
GISS.E2.H.CC		[0.48,0.59]	[0.015,0.019]	[-0.12,0.05]	[0.008,0.016]	[13.06,20.07]	[2.62,6.49]

Table 3: Table of parameter estimates. Two rows for each model: first are the estimates and second the 95% Credible Intervals for each parameter. Notation p is shorthand for the posterior probability of a particular model.

CGCM	Most likely model	β_S	σ_S	β_T	σ_T	ϕ_S	ϕ_T
GISS.E2.R.CC	M_S $(p = 0.99)$	0.52	0.026	-	-	-	-
GISS.E2.R.CC	M_{ST} $(p = 0.01)$	[0.48,0.55]	[0.015,0.018]	-	-	-	-
NorESM1.M	M_T $(p = 0.58)$	-	-	0.46	0.017	-	-
NorESM1.M	M_{ST} $(p = 0.42)$	-	-	[0.42,0.50]	[0.016,0.019]	-	-
GFDL.CM3	M_{ST} $(p = 1)$	0.42	0.030	-0.01	0.025	13.84	3.37
GFDL.CM3		[0.37,0.48]	[0.027,0.032]	[-0.08,0.07]	[0.019,0.032]	[8.48,23.86]	[2.27,4.77]
GFDL.ESM2M	M_{ST} $(p = 1)$	0.25	0.023	0.24	0.020	9.04	3.23
GFDL.ESM2M		[0.21,0.28]	[0.021,0.026]	[0.17,0.31]	[0.016,0.026]	[5.61,11.44]	[2.22,4.42]
CESM1.CAM5	M_{ST} $(p = 1)$	0.44	0.019	0.38	0.019	8.40	5.23
CESM1.CAM5		[0.36,0.53]	[0.016,0.021]	[0.27,0.53]	[0.015,0.022]	[4.87,14.60]	[3.06,8.07]
CESM1.WACCM	M_{ST} $(p = 1)$	0.22	0.019	0.35	0.014	43.50	55.50
CESM1.WACCM		[0.14,0.31]	[0.017,0.023]	[0.27,0.43]	[0.012,0.017]	[30.51,54.99]	[55.48,71.20]
HadGEM3-GC2	M_{ST} $(p = 1)$	0.36	0.019	0.55	0.015	12.80	26.76
HadGEM3-GC2		[0.25,0.47]	[0.014,0.024]	[0.47,0.65]	[0.013,0.017]	[7.82,22.91]	[17.95,43.10]
bcc.csm1.1.m	M_S $(p = 0.95)$	0.62	0.013	-	-	-	-
bcc.csm1.1.m	M_{ST} $(p = 0.05)$	[0.59,0.66]	[0.012,0.014]	-	-	-	-
CMCC.CM	M_{ST} $(p = 1)$	0.72	0.018	-0.30	0.026	59.29	6.71
CMCC.CM		[0.70,0.75]	[0.017,0.020]	[-0.44,-0.15]	[0.019,0.037]	[52.25,66.39]	[4.10,9.99]
CanESM2	M_{ST} $(p = 1)$	0.70	0.025	0.67	0.028	16.75	4.01
CanESM2		[0.66,0.73]	[0.023,0.027]	[0.55,0.77]	[0.023,0.033]	[13.97,19.44]	[3.24,5.02]
ACCESS1.0	M_{ST} $(p = 1)$	0.36	0.036	0.28	0.031	15.63	4.42
ACCESS1.0		[0.32,0.41]	[0.033,0.039]	[0.18,0.38]	[0.025,0.038]	[10.54,20.59]	[2.76,6.15]
CNRM.CM5	M_{ST} $(p = 0.94)$	0.76	0.026	0.56	0.030	93.93	7.04
CNRM.CM5	M_S $(p = 0.06)$	[0.74,0.79]	[0.025,0.027]	[0.25,0.78]	[0.022,0.041]	[83.80,105.33]	[3.57,11.55]
CSIRO.Mk3.6.0	M_{ST} $(p = 1)$	0.50	0.023	0.27	0.026	28.90	5.30
CSIRO.Mk3.6.0		[0.47,0.53]	[0.021,0.025]	[0.17,0.36]	[0.021,0.031]	[15.77,21.14]	[3.28,7.52]
EC.EARTH	M_S $(p = 0.52)$	0.54	0.022	-	-	-	-
EC.EARTH	M_{ST} $(p = 0.48)$	[0.50,0.58]	[0.021,0.024]	-	-	-	-
IPSL.CM5A.LR	M_{ST} $(p = 1)$	0.76	0.025	0.17	0.044	99.12	8.88
IPSL.CM5A.LR		[0.74,0.78]	[0.024,0.026]	[0.00,0.32]	[0.035,0.052]	[84.06,116.20]	[6.55,11.45]

Table 4: Table of parameter estimates. Two rows for each model: first are the estimates and second the 95% Credible Intervals for each parameter. Notation p is shorthand for the posterior probability of a particular model.

CGCM	Most likely model	β_S	σ_S	β_T	σ_T	ϕ_S	ϕ_T
IPSL.CM5B.LR	M_{ST} $(p = 0.98)$	0.76	0.026	0.67	0.051	25.76	5.05
IPSL.CM5B.LR	M_S $(p = 0.02)$	[0.75,0.83]	[0.023,0.030]	[0.33,1.04]	[0.035,0.068]	[17.68,34.34]	[2.89,7.43]
FGOALS.s2	M_{ST} $(p = 1)$	0.75	0.014	0.60	0.017	22.70	4.59
FGOALS.s2		[0.70,0.79]	[0.013,0.016]	[0.45,0.76]	[0.013,0.021]	[18.51,27.64]	[2.99,6.46]
HadGEM2.ES	M_{ST} $(p = 1)$	0.79	0.028	0.40	0.031	9.78	10.09
HadGEM2.ES		[0.72,0.86]	[0.026,0.032]	[0.34,0.47]	[0.028,0.034]	[8.04,11.81]	[7.95,12.13]
MPI.ESM.MR	M_{ST} $(p = 1)$	0.69	0.021	0.64	0.020	12.96	8.83
MPI.ESM.MR		[0.64,0.74]	[0.019,0.022]	[0.57,0.72]	[0.017,0.022]	[10.71,15.29]	[6.79,11.23]
MRI.CCGCM3	M_{ST} $(p = 1)$	0.72	0.022	0.56	0.031	16.31	2.95
MRI.CCGCM3		[0.67,0.77]	[0.020,0.024]	[0.32,0.77]	[0.023,0.039]	[9.71,22.56]	[2.10,3.98]
GISS.E2.H	M_{ST} $(p = 1)$	0.41	0.031	0.28	0.029	4.17	4.18
GISS.E2.H		[0.31,0.52]	[0.026,0.036]	[0.21,0.36]	[0.023,0.034]	[1.90,7.87]	[2.06,7.19]
GISS.E2.R	M_{ST} $(p = 0.89)$	0.53	0.015	0.32	0.016	56.25	2.48
GISS.E2.R	M_S $(p = 0.11)$	[0.50,0.55]	[0.014,0.016]	[0.09,0.55]	[0.004,0.029]	[46.46,62.67]	[1.29,4.17]
NorESM1.ME	M_{ST} $(p = 0.55)$	0.50	0.02	0.44	0.015	3.18	22.5
NorESM1.ME	M_T $(p = 0.45)$	[0.21,0.81]	[0.01,0.03]	[0.38,0.50]	[0.01,0.02]	[1.09,6.07]	[8.94,48.67]
GFDL.ESM2G	M_{ST} $(p = 0.98)$	0.57	0.020	0.26	0.028	70.66	3.77
GFDL.ESM2G	M_S $(p = 0.02)$	[0.55,0.58]	[0.018,0.021]	[-0.17,0.23]	[0.016,0.042]	[63.57,78.51]	[1.95,6.13]
CCSM4	M_{ST} $(p = 1)$	0.43	0.013	0.51	0.014	5.46	5.50
CCSM4		[0.38,0.48]	[0.012,0.014]	[0.45,0.58]	[0.013,0.015]	[4.37,6.75]	[4.37,6.92]
CESM1.FASTCHEM	M_{ST} $(p = 1)$	0.83	0.014	0.70	0.015	5.35	4.86
CESM1.FASTCHEM		[0.72,0.96]	[0.011,0.016]	[0.56,0.84]	[0.012,0.018]	[3.38,10.47]	[3.18,8.89]
CESM1.BGC	M_{ST} $(p = 1)$	0.48	0.012	0.63	0.017	4.45	3.44
CESM1.BGC		[0.42,0.55]	[0.010,0.014]	[0.51,0.75]	[0.014,0.019]	[3.39,5.86]	[2.54,4.45]
HadGEM3	M_{ST} $(p = 1)$	0.02	0.018	0.32	0.017	19.20	67.49
HadGEM3		[-0.02,0.06]	[0.016,0.021]	[0.30,0.34]	[0.016,0.019]	[14.69,24.46]	[54.71,81.14]

Table 5: Table of parameter estimates. Two rows for each model: first are the estimates and second the 95% Credible Intervals for each parameter. Notation p is shorthand for the posterior probability of a particular model.

Viterbi algorithm for HSMMs

The Viterbi algorithm aims to efficiently maximise $p(\boldsymbol{C}|\boldsymbol{\rho}, \boldsymbol{\theta})$ w.r.t. $\boldsymbol{C}$. This is performed recursively as described in Section 3.2 of Scott (2002). For HSMMs, this is not as straightforward since the HSMM can be viewed as an non-stationary HMM, i.e. one where the transition matrix is different for each time step. Fortunately, the forward and backward algorithms described in Economou et al. (2014) can be modified in such a way that this time varying transition matrix is computed explicitly and the Viterbi algorithm applied as in Scott (2002).

Firstly we note that due to conditional independence, maximizing $p(\boldsymbol{C}|\boldsymbol{\rho}, \boldsymbol{\theta})$ is the same as maximizing $p(\boldsymbol{C}, \boldsymbol{\rho}|\boldsymbol{\theta})$ w.r.t. $\boldsymbol{C}$. So the same forward algorithm used to compute the likelihood, can be instead used to maximise it. This can be done simply by modifying the algorithm on page 8 of Economou et al. (2014), specifically replacing $\gamma_{T_N}(j) + \sum_{u=2}^{T_{N-1}} \xi_u(j)$ with $\max_j \left(\gamma_{T_N}(j) + \sum_{u=2}^{T_{N-1}} \xi_u(j)\right)$. This will yield new vectors $\alpha_{T_N}(j)$, which can then be multiplied by ℓ_{T_N} to yield quantities $L_{T_N}(j)$ as defined in Scott (2002). The backward step can be followed exactly like in Section 3.2 of Scott (2002) where the transition matrix $q(i, j)$ is obtained by computing $a_{i,j,T_N} \ell_{T_N} F_j(N) / \ell_{T_{N-1}}$.