Proper orthogonal decomposition reduced-order model of the global oceans

Vassili Kitsios, Laurent Cordier, Terence O'kane

Proper orthogonal decomposition reduced-order model of the global oceans

Vassili Kitsios[1,2*], Laurent Cordier[3] and Terence J. O'Kane[4]

[1*]Environment, CSIRO, 107-121 Station Street, Aspendale, 3195, Victoria, Australia.
[2]Laboratory for Turbulence Research in Aerospace and Combustion, Department of Mechanical and Aerospace Engineering, Monash University, Clayton, 3800, Victoria, Australia.
[3]Institut Pprime, CNRS, Université de Poitiers, ENSMA, Département Fluides Thermique et Combustion, 86360, Futuroscope-Chasseneuil, France.
[4]Environment, CSIRO, Castray Esplanade, Battery Point, 7004, Tasmania, Australia.

*Corresponding author(s). E-mail(s): Vassili.Kitsios@csiro.au;

Abstract

A reduced-order model (ROM) of the global oceans is developed by projecting the hydrostatic Boussinesq equations of motion onto a proper orthogonal decomposition (POD) basis. Three-dimensional POD modes are calculated from the ocean fields of an ensemble climate reanalysis dataset. The coefficients in the POD ROM are calculated using a regression approach. The performance of various POD ROM configurations are assessed. Each configuration is derived from an alternate sea-water equation of state, linking the density and temperature fields. POD ROM variants incorporating an equation of state in which density is a quadratic function of temperature, are able to reproduce the statistics of the large-scale structures at a fraction of the computational cost required to numerically simulate this flow. Due to the speed and efficiency of calculation, such reduced-order models of the global geophysical system will enable researchers and policy makers to assess the physical risk for a broader range of potential future climate scenarios.

Keywords: reduced-order modelling, ocean, climate

1 Introduction

Geophysical turbulence is highly dimensional, nonlinear, and multi-scale. Scales of motion range from planetary waves with lengths in the order of $10,000$km to millimetre sized turbulence [1]. It is chaotic, whilst still comprising of large-scale three-dimensional coherent structures with significant temporal and spatial correlations [2, 3]. Our ability to understand and predict such physical phenomena benefits from reducing these systems to their most basic building blocks. As Einstein famously put it in his 1933 lecture, *"It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience"* [4]. This quote has commonly been paraphrased to state that a model should be as simple as possible but not simpler. One might consider reduced-order modelling as a mathematical representation of this statement. Here, we develop such reduced-order models (ROM) to simulate the global oceans.

There are numerous approaches to model reduction. In a Galerkin projection, continuous equations of motion are solved on a truncated basis. Such a basis serves as a collection of the aforementioned basic elements, or building blocks. There are many techniques to calculate such a basis tailored to the flow configuration of interest. Methods include, but are not limited to: linear stability (or normal) modes [5–7]; principle oscillation pattern type methods [8–11]; finite time variants of normal modes [12]; cyclic principle oscillation patterns [13]; covariant Lyapunov vectors [14]; and bred vectors [15]. Many of these methods can be considered as subsets of a more general convex coding framework [16]. The basis adopted in this study, is the proper orthogonal decomposition (POD), also referred to in other fields as empirical orthogonal functions, or a singular value decomposition.

A POD is a set of orthogonal basis functions for a given series of observations (or snapshots). These modes can be used to describe the unsteady motions of turbulent flows [17]. The POD basis is designed to maximise a specified non-negative norm (e.g. kinetic energy). By definition, this norm decays monotonically for each subsequent mode. To develop a POD ROM, one projects appropriate equations of motion onto the POD basis, producing a set of ordinary differential equations (ODEs). These ODEs can be solved at a fraction of the computational cost used to generate the original data set. One can calculate the coefficients in the ODEs from their analytical expressions [18, 19] or using regression approaches [20, 21]. These approaches are designed to evolve the instantaneous flow field forward in time. Alternate Bayesian [22] and statistical dynamical closure methods [23] simulate properties of the probability distribution function. In this study, we will develop the first ever POD ROM of the global oceans, as represented within a global reanalysis dataset.

In general, reanalyses are generated using some form of data assimilation to modify imperfect simulations of reality with a series of partial and potentially noisy measurements [24]. This results in a better representation of the true system state than could be achieved with measurements or simulations alone. Numerical simulations of the global oceans (and atmosphere) are undertaken by codes referred to as general circulation models (GCMs). Freely running GCMs can possess significant biases, due in part to the: unresolved scales of motion; parameterisation of unresolved physical

processes; and the artificial dissipation introduced by spatial discretisation and time stepping schemes [25]. Reanalyses mitigates against such biases and provide the best possible estimate of the historical climate.

However, reanalyses are limited to the period over which sufficient historical observations are available. The atmosphere has been well observed since the introduction of satellite monitoring in the 1970s. Satellites have also provided sufficient observations for the ocean surface temperatures, but not within the ocean volume. The global ocean interior has arguably only been well observed since the mid to late 2000s (see figures 1 and 2 of [26]), coinciding with the increased density of the in-situ ocean monitoring network [27]. Fortunately, the Climate Analysis Forecast Ensemble (CAFE) reanalysis, denoted by CAFE-60 [26], provides 96 realisations of the Earth each month. All realisations satisfy the equations of motion, and the partial observations of the Earth system. CAFE-60 is the dataset adopted herein. It has been shown to be consistent with other world-class data sets on multiple fronts, including the representation of large scale climate phenomena (e.g. the El Niño / La Niña cycle) [28].

The manuscript is organised as follows. The CAFE-60 reanalysis is first characterised in section 2. The snapshot POD method is presented in section 3, and applied to the velocity and temperature fields. The POD ROM dynamical system is derived in section 4, with the hydrostatic Boussinesq equations of motion for the global ocean projected onto the POD modes. A derivation of the oceanic equations of motion can be found in appendix A. The most general equation of state adopted herein, is one in which the density of seawater is a quadratic function of temperature. We also assess equations of state in which density is linearly dependent upon temperature, and also one with no temperature dependence. Each of these assumed equations of state yield different sets of coefficients in the POD ROM dynamical system. In section 5, linear regression is used to calculate the POD ROM coefficients associated with these equations of state, as well as some additional variants. Further details on the regression approach is presented in appendix B. We use all of the CAFE-60 ensemble members to learn the POD ROM coefficients, which provides us with a factor of 96 times more samples than one would have if only one state estimate was available per time instant. The evolution and assessment of the optimal POD ROM system is presented in section 6. Finally, concluding remarks are made in section 7.

2 Reanalysis and flow characterisation

In CAFE-60 [26, 28], numerical simulations and real-world observations are fused together via the ensemble transform Kalman filter algorithm [29, 30]. In essence, this algorithm sums together the simulated and observed data inversely weighted by their respective uncertainties. The CAFE system manages an ensemble of 96 simultaneous numerical simulations of the global climate starting from different initial conditions. The spatio-temporally varying model uncertainty is quantified by the covariance matrix calculated across the 96 ensemble members. The adopted GCM solves for the coupled global atmosphere, ocean, sea-ice, land and bio-geo-chemical system [31]. The correction of all prognostic variables within the GCM are determined on the basis of a comprehensive network of global real world observations, with uncertainties prescribed

for each observation class (e.g. satellites, floats). This approach has also been used to simultaneously estimate both prognostic variables and model parameters [32]. CAFE-60 provides three-dimensional fields for the entire ensemble every month from 1960 to 2020. It is effectively a spatio-temporally varying sampled probability distribution function of the global climate.

Whilst CAFE-60 provides the prognostic variables pertaining to the atmosphere, ocean, sea-ice, land and bio-geo-chemistry, in this study we require only the ocean variables. We analyse the monthly averaged ocean fields over the period from January 2010 to December 2020, which are a function of time t, and space $\mathbf{x} = (x, y, z)$. The spatial coordinates are attached to the rotating Earth, where x is positive pointing east (longitudinal direction), y is positive pointing north (latitudinal or meridional direction), and z is positive pointing ~~up~~upward normal to the surface of the Earth. The Earth's surface is located at $z = 0$, with ocean depth defined as $-z$. The zonal (south to north), meridional (west to east) and vertical velocity components are denoted by u, v and w, respectively. The ocean grid has 50 vertical levels, with grid spacings of 10m up to a depth of 200m. The vertical grid spacings then increase as they approach the latitudinally and longitudinally dependent ocean floor, which is nowhere deeper than 6km. The grid is unstructured in the horizontal plane, however, it nominally has a longitudinal resolution of 1°, with the latitudinal resolution finer in specific regions. In the generation of the data, the time step size of the ocean model component was 1 hour. The data has been output as monthly averages, hence the time interval between the samples is one month.

To facilitate the following discussion, ~~for a given ensemble member (e),~~ we define the state vector $\mathbf{q}_e = (\mathbf{u}_e, T_e)$, comprising of the temperature (T_e), and horizontal velocity vector $\mathbf{u}_e = (u_e, v_e)$ with eastward zonal (u_e), and northward meridional (v_e) velocity components. The subscript $_e$ on each of the state variables refers to the ensemble member index. In the CAFAE-60 dataset e ranges from 1 to 96, inclusive. ~~we~~Note, we focus on only the horizontal velocity components, since in the hydrostatic approximation, the vertical velocity component is a diagnostic field as opposed to a prognostic one - see appendix A.

The boundary conditions of the physical system are both periodic (e.g. solar radiation) and aperiodic (e.g. growing greenhouse gas concentration) in nature. This gives rise to trends, and variability on annual and inter-annual timescales in the monthly averaged fields. Since we are only considering the most recent decade, the trend component is negligible, and can safely be ignored. The flow is then decomposed using the triple decomposition of [33] such that

$$\mathbf{q}_e(\mathbf{x}, t) = \bar{\mathbf{q}}_e(\mathbf{x}) + \widetilde{\mathbf{q}}_e(\mathbf{x}, t) + \mathbf{q}'_e(\mathbf{x}, t) \ , \tag{1}$$

where $\bar{\mathbf{q}}_e(\mathbf{x})$ is the time average, and $\widetilde{\mathbf{q}}_e(\mathbf{x}, t)$ is the seasonal component with a 1-year phase period and time mean of zero. The remaining $\mathbf{q}'_e(\mathbf{x}, t)$ term, represents the anomalous inter-annual fluctuations about the seasonal cycle. We also define $\breve{\mathbf{q}}_e(\mathbf{x}, t) = \bar{\mathbf{q}}_e(\mathbf{x}) + \widetilde{\mathbf{q}}_e(\mathbf{x}, t)$, as the phase (or climatological) component. The phase angles are the twelve calendar months

$$\vartheta \in [\text{January}, \ldots, \text{December}] \equiv [\text{Jan}, \ldots, \text{Dec}] \equiv [1, \ldots, 12] \ . \tag{2}$$

The climatological component is reconstructed according to

$$\breve{\mathbf{q}}_e(\mathbf{x},t) \;=\; \bar{\mathbf{q}}_e(\mathbf{x}) + \sum_{\vartheta=\mathrm{Jan}}^{\mathrm{Dec}} \widetilde{\mathbf{q}}_e^{\,\vartheta}(\mathbf{x})\; d^{\vartheta}(t) \equiv \sum_{\vartheta=\mathrm{Jan}}^{\mathrm{Dec}} \breve{\mathbf{q}}_e^{\,\vartheta}(\mathbf{x})\; d^{\vartheta}(t)\;, \tag{3}$$

where $d^{\vartheta}(t)$ is a time series equal to 1 within the associated month ϑ, and zero otherwise, with the property

$$\sum_{\vartheta=\mathrm{Jan}}^{\mathrm{Dec}} d^{\vartheta}(t) = 1 \quad , \tag{4}$$

for all time t. The term $\widetilde{\mathbf{q}}_e^{\,\vartheta}(\mathbf{x})$ is the average deviation from the mean of season ϑ. The associated phase averaged field is calculated in the usual way according to

$$\breve{\mathbf{q}}_e^{\,\vartheta}(\mathbf{x}) = \bar{\mathbf{q}}_e(\mathbf{x}) + \widetilde{\mathbf{q}}_e^{\,\vartheta}(\mathbf{x}) = \sum_{\iota=1}^{N_{\mathrm{years}}} \mathbf{q}_e(\mathbf{x}, t_0 + \vartheta + \iota N_{\vartheta}) \quad , \tag{5}$$

for each phase angle ϑ, where t_0 is the time of the first data instance, N_{years} the number of years in the data set, and $N_{\vartheta} = 12$ being the number of months per year (i.e. number of phase angles). The ensemble average is the best estimate of the system state given by

$$\mathbf{q}(\mathbf{x},t) \;=\; \frac{1}{N_{\mathrm{ens}}} \sum_{e=1}^{N_{\mathrm{ens}}} \mathbf{q}_e(\mathbf{x},t)\;, \tag{6}$$

where the number of ensemble members $N_{\mathrm{ens}} = 96$. One can also take the ensemble average of the individual components within the triple decomposition.

To characterise the flow, properties in the latitude / longitude plane at the sea surface are illustrated in figure 1. The rows from top to bottom illustrate statistics for the zonal (west to east) velocity, meridional (south to north) velocity and temperature fields. The columns from left to right illustrate the mean, standard deviation of the seasonal cycle, and standard deviation of the fluctuations about the seasonal cycle. To facilitate a direct comparison of the zonal velocity standard deviations, the same colour bars are adopted for the seasonal component in figure 1(b) and the anomalous fluctuations in figure 1(c). A common colour bar is also used for the associated standard deviations of the meridional velocity in figures 1(e) and 1(f). Typical of the Earth system, the variability in the seasonal cycle is larger than the inter-annual variability in most locations, and particularly so in the tropics. This is a result of the seasonal changes in the solar forcing, due to the Earth's inclined axis of rotation, and its elliptical orbit around the Sun. The seasonal variability in the fluid fields arises due to shifts in location of the mean structures. The Southern Ocean is perhaps an exception, where many of the mean velocity structures are persistent, and have a lesser dependence upon the seasons, as quantified by the lower seasonal standard deviations in

this region. This lack of seasonality is due to the ocean currents in this region being steered by the continental geometry and ocean floor topography. Whilst we present here the mean and standard deviations of the flow field, the velocity fields are in fact non-Gaussian in time, with a spatially dependent skewness.

The mean temperature field has a strong dependence upon latitude as illustrated in figure 1(g). It is warmer in the tropics and cooler in the polar regions due to the additional distance the solar radiation must travel to reach the higher latitudes. The seasonal variability in temperature is significantly greater than its inter-annual variability. Notice the change in colour bar limits in figures 1(g) and 1(h). This difference in variability is more stark than what was previously observed for the velocity fields. Unlike the velocity field, the temperature field in the Southern Ocean does have significant seasonal variability. This is due to the strong influence the cyclical solar forcing has on ocean stratification in this region. The Pacific ocean is the zone of greatest variability in the fluctuating temperature field in figure 1(h). It is coincident with the principle location of the El Niño Southern Oscillation, with canonical phases El Niño and La Niña [34]. The POD presented in the following section decomposes the fluctuating (or anomalous) velocity and temperature fields to further characterise the dynamics. The temperature fields are also non-Gaussian in time, with the spatially dependent non-zero skewness.

3 Proper Orthogonal Decompositions

The snapshot POD method [35] is utilised in this study as it is computationally more efficient when the spatial resolution exceeds the temporal resolution, which is in case here. As required for the POD ROM to follow, we calculate a POD for the fluctuating components of the ensemble averaged fields. The ensemble averaged state vector is decomposed into its mean, seasonal and fluctuating components according to

$$\mathbf{q}(\mathbf{x},t) = \bar{\mathbf{q}}(\mathbf{x}) + \widetilde{\mathbf{q}}(\mathbf{x},t) + \mathbf{q}'(\mathbf{x},t) \quad . \tag{7}$$

One POD is calculated for the anomalous horizontal velocity vector field ($\mathbf{u}'$), and another for the anomalous temperature scalar field (T'). It is perhaps non-standard in the fluid mechanics literature to calculate POD modes on the fluctuations about a cycle, as opposed to the fluctuations about a static mean. This is perhaps because in engineering applications, one is interested in how the period of $\widetilde{\mathbf{q}}(\mathbf{x},t)$, might change under different Reynolds numbers and other flow parameters. In the present geophysical application, however, $\widetilde{\mathbf{q}}(\mathbf{x},t)$ is the seasonal cycle, with a phase period fixed by the Earth's orbit. In this instance, one is primarily interested in understanding the anomalies and making predictions more skilful than prescribing the repeating seasonal cycle. Given this motivation, the snapshot POD method is applied to the fluctuations about the phase average.

The POD of the anomalous horizontal velocity field requires the solution of the following eigenvalue problem

$$\mathbf{E}\mathbf{a}^{(n)} = \Lambda_u^{(n)}\mathbf{a}^{(n)} \quad , \tag{8}$$

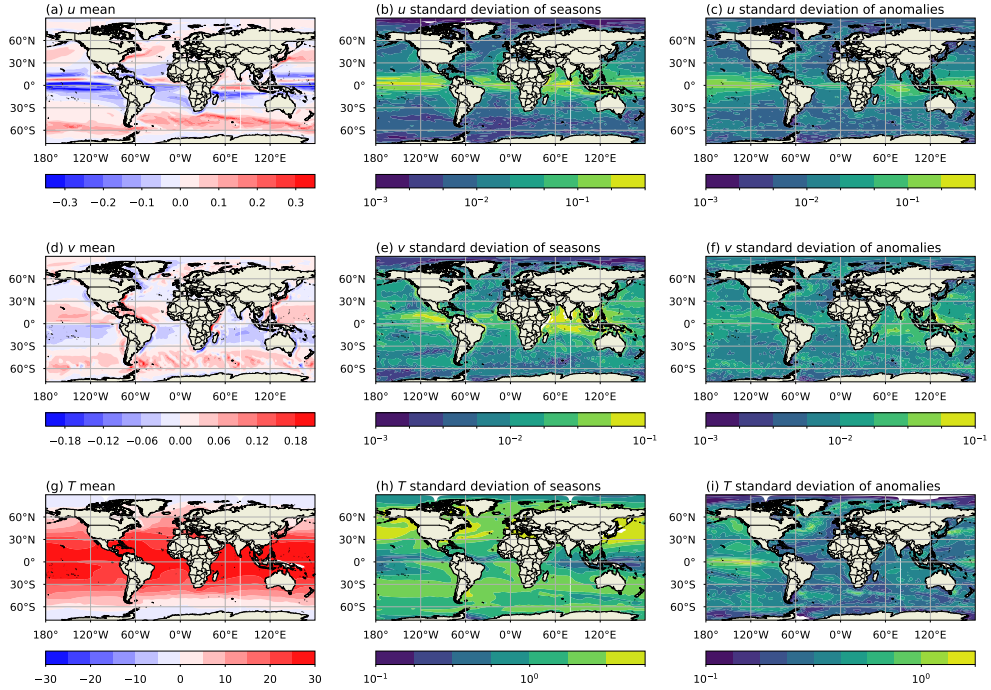

Fig. 1 Statistical properties at the sea surface. Time mean (first column), standard deviation of the seasonal cycle (second column), and standard deviation of the fluctuations (third column) for the zonal velocity (first row in ms^{-1}), meridional velocity (second row in ms^{-1}) and temperature (third row in $^{\circ}C$) fields.

where the elements of the covariance matrix $\mathbf{E}$ are given by

$$E_{ki} = \frac{1}{N_s} \left\langle \mathbf{u}'(\mathbf{x}, t_k), \mathbf{u}'(\mathbf{x}, t_i) \right\rangle_u \quad \text{for} \quad 0 < k, i < N_s - 1 \quad , \tag{9}$$

with N_s the number of snapshots. The inner product is defined as

$$\left\langle \mathbf{u}'(\mathbf{x}, t_k), \mathbf{u}'(\mathbf{x}, t_i) \right\rangle_u = \int_V \mathbf{u}'(\mathbf{x}, t_k) \cdot \mathbf{u}'(\mathbf{x}, t_i) \, dV \quad , \tag{10}$$

where V is the integration volume. $\mathbf{E}$ is a positive symmetric matrix of leading dimension N_s. The matrix $\mathbf{E}$, therefore, has N_s non-negative real eigenvalues $\Lambda_u^{(n)}$, and associated eigenvectors $\mathbf{a}^{(n)}$. This inner product is representative of the kinetic energy in the anomalies of the horizontal velocity components throughout the entire ocean volume, and is also required for the POD ROM to follow.

The elements in the eigenvector $\mathbf{a}^{(n)}$, are instances in time of the temporal POD mode $a^{(n)}(t)$. The temporal modes are normalised such that

$$\overline{a^{(n)}(t)\ a^{(m)}(t)} = \frac{1}{N_s}\sum_{k=1}^{N_s} a^{(n)}(t_k)\ a^{(m)}(t_k) = \Lambda_u^{(n)}\delta_{nm} \quad , \tag{11}$$

which ensures the amplitude is proportional to the kinetic energy in the mode, and δ_{nm} is the Kronecker delta function. The spatial modes $\boldsymbol{\mathcal{U}}^{(n)}(\mathbf{x}) \equiv (\mathcal{U}^{(n)}(\mathbf{x}), \mathcal{V}^{(n)}(\mathbf{x}))$, are then calculated via

$$\boldsymbol{\mathcal{U}}^{(n)}(\mathbf{x}) = \frac{1}{N_s\Lambda_u^{(n)}}\sum_{k=1}^{N_s} a^{(n)}(t_k)\ \mathbf{u}'(\mathbf{x},t_k)\ , \tag{12}$$

which by definition are also orthogonal such that

$$\left\langle \boldsymbol{\mathcal{U}}^{(n)}(\mathbf{x}), \boldsymbol{\mathcal{U}}^{(m)}(\mathbf{x}) \right\rangle_u = \delta_{nm} \quad . \tag{13}$$

The horizontal velocity vector field can then be reconstructed by

$$\begin{aligned} \mathbf{u}(\mathbf{x},t) &= \bar{\mathbf{u}}(\mathbf{x}) + \widetilde{\mathbf{u}}(\mathbf{x},t) + \mathbf{u}'(\mathbf{x},t) \\ &\simeq \bar{\mathbf{u}}(\mathbf{x}) + \sum_{\vartheta=\mathrm{Jan}}^{\mathrm{Dec}} d^{\vartheta}(t)\ \widetilde{\mathbf{u}}^{\vartheta}(\mathbf{x}) + \sum_{n=1}^{N_u} a^{(n)}(t)\ \boldsymbol{\mathcal{U}}^{(n)}(\mathbf{x}) \quad . \end{aligned} \tag{14}$$

The reconstruction is exact when $N_u = N_s$. The equations of motion are projected onto this decomposition in the POD ROM section to follow.

Following the same procedure we develop the temperature decomposition

$$\begin{aligned} T(\mathbf{x},t) &= \bar{T}(\mathbf{x}) + \widetilde{T}(\mathbf{x},t) + T'(\mathbf{x},t) \\ &\simeq \bar{T}(\mathbf{x}) + \sum_{\vartheta=\mathrm{Jan}}^{\mathrm{Dec}} d^{\vartheta}(t)\ \widetilde{T}^{\vartheta}(\mathbf{x}) + \sum_{n=1}^{N_T} b^{(n)}(t)\ \mathcal{T}^{(n)}(\mathbf{x}) \quad , \end{aligned} \tag{15}$$

where $b^{(n)}(t)$ is the n-th temporal POD mode, and N_T the number of modes retained in the reconstruction. The inner product associated with the scalar temperature field is given by

$$\langle T'(\mathbf{x},t_k), T'(\mathbf{x},t_i)\rangle_T = \int_V T'(\mathbf{x},t_k)\ T'(\mathbf{x},t_i)\ dV \quad . \tag{16}$$

It has analogous orthogonality properties of $\left\langle \mathcal{T}^{(n)}(\mathbf{x}), \mathcal{T}^{(m)}(\mathbf{x}) \right\rangle_T = \delta_{nm}$ and $\overline{b^{(n)}(t)\ b^{(m)}(t)} = \Lambda_T^{(n)}\delta_{nm}$, where $\Lambda_T^{(n)}$ is the variance of each mode. This decomposition will contribute to the source term in the POD ROM presented in the following section.

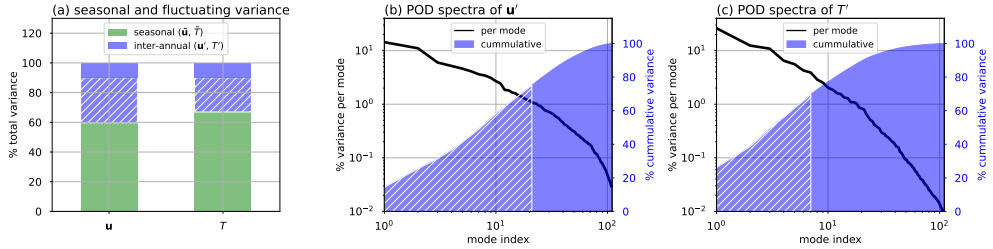

Fig. 2 Decomposition of variability in the horizontal velocity ($\mathbf{u}$) and temperature (T) fields. (a) Contribution of seasonal ($\tilde{\mathbf{u}}, \tilde{T}$) and anomalous variability ($\mathbf{u}', T'$), with white hatched regions indicating the amount of anomalous variance required to reach 90% of total variance. (b) POD spectra of $\mathbf{u}'$. (c) POD spectra of T'. White hatched regions in (b) and (c) indicate the number of POD modes required to reach 90% of total variance.

We now describe the distribution of variability in the velocity and temperature fields. Figure 2(a) illustrates the breakdown of the globally integrated variability between the seasonal (green) and anomalous (blue) components. As discussed in the previous section, the seasonal component is the dominant source of variability in the climate system. Here the seasonal component contributes 60% of the total horizontal velocity field variability and 67% of the total temperature variability. The white hatched zone in these bar graphs indicate the additional amount of variability from the fluctuating component required to reach 90% of the total variability in the system. The variability associated with each mode in the velocity and temperature field POD are illustrated in figures 2(b) and 2(c), respectively. The black lines illustrate the variance per mode as a percentage of the total anomalous variance in the decomposition. The temperature POD has a greater concentration of variability in the first few modes and a steeper decay of variance with mode index as compared to the velocity POD spectra. The blue shaded region illustrates the cumulative variance again as a percentage of the total anomalous variance. Consistent with figure 2(a), the white hatched region indicates the amount of anomalous variance required, which in addition to the seasonal component, captures 90% of the total variance. For the velocity field 20 POD modes are required, whilst for the temperature field only 6 POD modes are needed.

These six most energetic temperature POD modes are illustrated in figure 3. The solid lines in figure 3(a) are the temporal modes $b^{(n)}(t)$. The multi-year time scale of the first two temporal POD modes are consistent with the nominal oscillation period between El Niño and La Niña states. The large amplitudes for modes $b^{(1)}$ and $b^{(2)}$ in the beginning of 2016 coincides with a large El Niño event. This particular event is characterised in more detail in figure 1 of [36]. The spatial patterns in the tropical Pacific ocean of the first two modes in figure 3(b) and figure 3(c) are also indicative of the variability associated with the El Niño Southern Oscillation. The higher order modes have shorter time scales, and comparatively more variability distributed in the higher latitude regions. We can also use this decomposition to determine the relative variability over the ensemble of climates. The shaded envelope in figure 3(a) is the range of coefficients across all 96 ensemble members. This is calculated by applying the inner product (16) between spatial mode $\mathcal{T}^{(n)}(\mathbf{x})$ and each temperature field $T_e(\mathbf{x}, t)$,

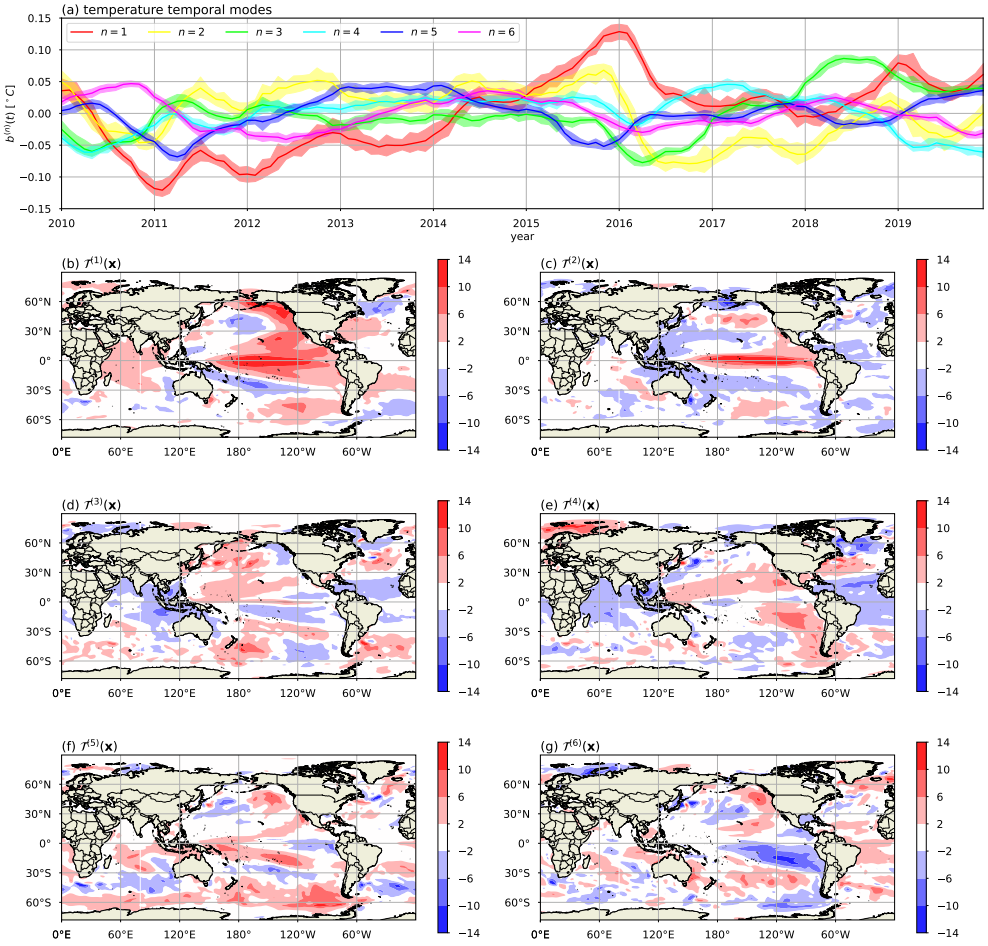

Fig. 3 POD of anomalous temperature variability. (a) Temporal modes of six most energetic modes, $b^{(n)}(t)$ (solid coloured lines), with range of coefficients for each mode across the 96 member ensemble (shaded coloured envelope), all in units of $^{\circ}C$. Associated dimensionless spatial modes: (b) $\mathcal{T}^{(1)}(\mathbf{x})$; (c) $\mathcal{T}^{(2)}(\mathbf{x})$; (d) $\mathcal{T}^{(3)}(\mathbf{x})$; (e) $\mathcal{T}^{(4)}(\mathbf{x})$; (f) $\mathcal{T}^{(5)}(\mathbf{x})$; and (g) $\mathcal{T}^{(6)}(\mathbf{x})$.

for all time t and all ensemble members e. The fact that these envelopes tightly follow the temporal modes, indicates that the variability of the ensemble mean over time, is greater than the uncertainty across the ensemble at a given instant in time.

The horizontal velocity field POD is illustrated in figure 4. The temporal modes in figure 4(a) have shorter dominant time scales as compared to those of the temperature POD. The spatial patterns of the velocity components are also of smaller scale as compared to the temperature modes. Note, the colour bars for the zonal velocity in left column are all the same. Likewise, the colour bars for the meridional velocity component in the right column are all the same. Moving from the top row of maps to the bottom, the mode index increases, and the variability is progressively distributed

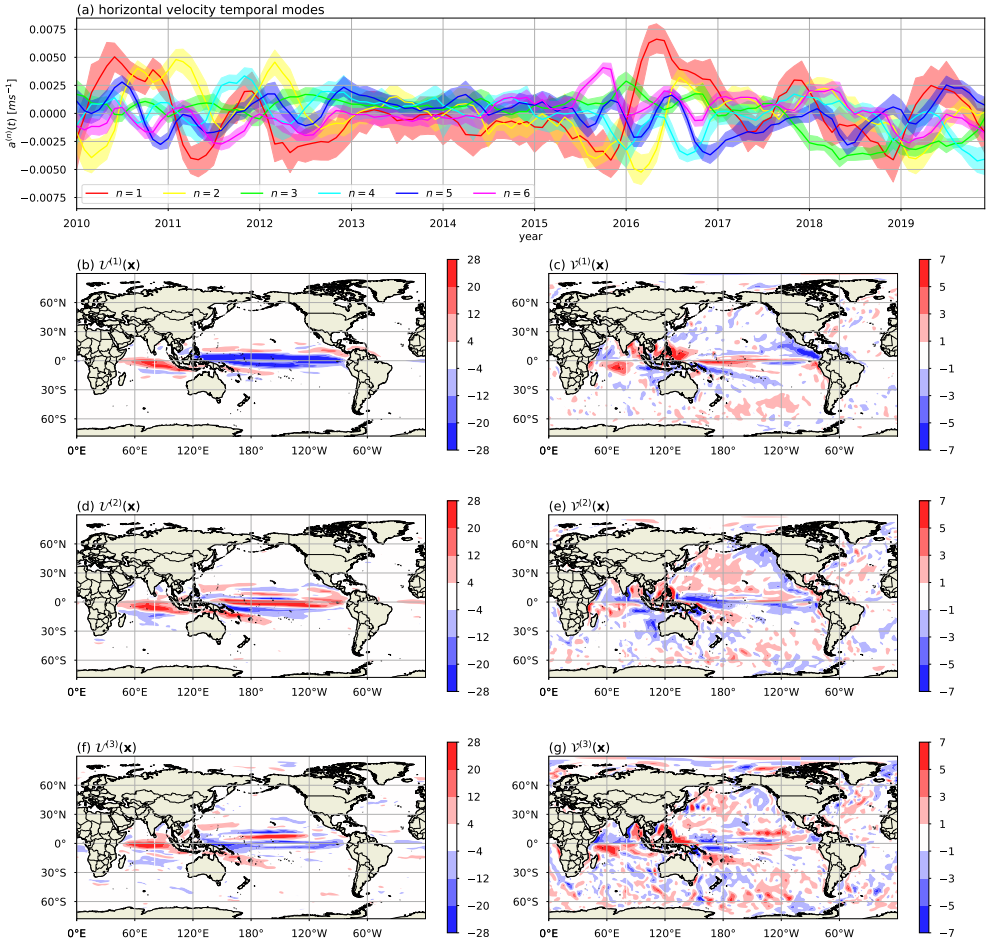

Fig. 4 POD of anomalous horizontal velocity variability. (a) Temporal modes of six most energetic modes, $a^{(n)}(t)$ (solid coloured lines), with range of contributions of each mode across the 96 member ensemble (shaded coloured envelope), all in units of ms^{-1}. Zonal and meridional components of the dimensionless spatial modes, for the three most energetic: (b) $\mathcal{U}^{(1)}(\mathbf{x})$; (c) $\mathcal{V}^{(1)}(\mathbf{x})$; (d) $\mathcal{U}^{(2)}(\mathbf{x})$; (e) $\mathcal{V}^{(2)}(\mathbf{x})$; (f) $\mathcal{U}^{(3)}(\mathbf{x})$; and (g) $\mathcal{V}^{(3)}(\mathbf{x})$.

away from the tropical regions and toward the higher latitudes. This is most evident in the meridional velocity. Returning to the temporal POD modes in figure 4(a), the width of the shaded range across the ensemble for these velocity modes is relatively larger than that observed for the temperature modes. More precisely, the variability across the ensemble relative to the variability of the ensemble mean in time, is larger for the temporal velocity POD modes as compared to the temperature ones. Note, in CAFE-60 ocean temperatures are directly observed via satellite measurements and floats. The velocity field on the other hand is not directly observed, and rather inferred

via its dynamical relationship to temperature and other observed quantities. Consequently there is a greater diversity of velocity fields that satisfy both the equations of motion and available observations. The velocity and temperature POD provide the basis for our ROM to follow.

4 Derivation of POD ROM dynamical system

Our POD ROM is based on the hydrostatic Boussinesq equations of motion for the global ocean. As derived in appendix A this system of equations can be written in the compact vector form

$$\frac{\partial \mathbf{u}}{\partial t} = \boldsymbol{\mathcal{N}}(\mathbf{u},\mathbf{u}) + \boldsymbol{\mathcal{L}}(\mathbf{u}) + \boldsymbol{\mathcal{M}}(T,T) + \boldsymbol{\mathcal{S}}(T) \text{ , where} \tag{17}$$

$$\boldsymbol{\mathcal{N}}(\mathbf{u},\mathbf{v}) = -\mathbf{u}\cdot\boldsymbol{\nabla}_h\mathbf{v} + \int_z^0 \boldsymbol{\nabla}_h\cdot\mathbf{u}\; d\underline{z}\; \frac{\partial \mathbf{v}}{\partial z} \text{ ,} \tag{18}$$

$$\boldsymbol{\mathcal{L}}(\mathbf{u}) = f\left[(\mathbf{e}_2\cdot\mathbf{u})\mathbf{e}_1 - (\mathbf{e}_1\cdot\mathbf{u})\mathbf{e}_2\right] \text{ ,} \tag{19}$$

$$\boldsymbol{\mathcal{M}}(T,\Upsilon) = -\frac{g}{\rho_0}\int_z^0 2\gamma_2 T\boldsymbol{\nabla}_h\Upsilon d\underline{z} \text{ ,} \tag{20}$$

$$\boldsymbol{\mathcal{S}}(T) = -\frac{g}{\rho_0}\int_z^0 (\gamma_1\underline{z} - 2\gamma_2 T_{\text{ref}})\boldsymbol{\nabla}_h T d\underline{z} \text{ ,} \tag{21}$$

and $\mathbf{u} = (u, v)$ as defined previously, with $\boldsymbol{\nabla}_h$ the horizontal derivative operator. Here $\underline{z}$ is used to distinguish the terms in the integrand from the lower limit in the integral. Note, z is the lower limit in the integral because the ocean surface is at $z = 0$, and since z is positive upward, then the ocean depths all have negative values of z. The unit vectors $\mathbf{e}_1 = (1, 0)$ and $\mathbf{e}_2 = (0, 1)$, the gravity $g = 9.81\text{ms}^{-2}$, and the Coriolis parameter $f = 2\Omega\sin\phi$, with ϕ the latitude, and $\Omega = 7.292\times10^{-5}\text{s}^{-1}$ the angular velocity magnitude of the Earth. The nominal ocean density is $\rho_0 \approx 1.04$ g cm^{-3}. ~~The vector $\mathbf{v}$ is of the same type as $\mathbf{u}$, and has been introduced here to write $\boldsymbol{\mathcal{N}}$ in the general manner required below.~~ $\boldsymbol{\mathcal{N}}$ is a nonlinear operator. For the purposes of defining (17) one could express $\boldsymbol{\mathcal{N}}$ as simply a function of $\mathbf{u}$. However, here we define $\boldsymbol{\mathcal{N}}$ as a function of the two vector quantities of equivalent type, $\mathbf{u}$ and $\mathbf{v}$, which is required to define the POD ROM coefficients below. Likewise, the scalar Υ is of the same type as T, and introduced to write $\boldsymbol{\mathcal{M}}$ in a sufficiently general way. The coefficients γ_1, γ_2, and T_{ref} are associated with the non-linear equation of state for seawater, linking the temperature field to density - see appendix A. The viscous term would ordinarily appear in the $\boldsymbol{\mathcal{L}}(\mathbf{u})$ term, but is negligible for the global ocean at the scales resolved in the current reanalysis.

We produce a POD ROM of this system by applying the horizontal velocity field inner product (10), between the n-th velocity spatial POD mode, $\boldsymbol{\mathcal{U}}^{(n)}$, and the equations of motion in (17), such that

$$\left\langle \boldsymbol{\mathcal{U}}^{(n)}, \frac{\partial \mathbf{u}}{\partial t}\right\rangle_u = \left\langle \boldsymbol{\mathcal{U}}^{(n)}, \boldsymbol{\mathcal{N}}(\mathbf{u},\mathbf{u}) + \boldsymbol{\mathcal{L}}(\mathbf{u}) + \boldsymbol{\mathcal{M}}(T,T) + \boldsymbol{\mathcal{S}}(T)\right\rangle_u \text{ .} \tag{22}$$

The POD ROM is derived by substituting in the velocity and temperature decompositions of (14) and (15), into (22). Expanding the terms and applying the orthogonality properties, results in the ordinary differential equations (ODE)

$$\begin{aligned}\dot{\hat{a}}^{(n)}(t) \;=\; & \sum_{\vartheta=\mathrm{Jan}}^{\mathrm{Dec}} \left[D_n^{\vartheta} \dot{d}^{\vartheta}(t) + C_n^{\vartheta} d^{\vartheta}(t) + \sum_{\vartheta'=\mathrm{Jan}}^{\mathrm{Dec}} F_n^{\vartheta\vartheta'} d^{\vartheta}(t) d^{\vartheta'}(t) \right] \\ & + \sum_{\vartheta=\mathrm{Jan}}^{\mathrm{Dec}} \left[\sum_{m=1}^{N_u} L_{nm}^{\vartheta} \hat{a}^{(m)}(t) d^{\vartheta}(t) + \sum_{m=1}^{N_T} Z_{nm}^{\vartheta} b^{(m)}(t) \right] \\ & + \sum_{m=1}^{N_u} \sum_{k=1}^{N_u} Q_{nmk} \hat{a}^{(m)}(t)\; \hat{a}^{(k)}(t) + \sum_{m=1}^{N_T} \sum_{k=1}^{N_T} R_{nmk} b^{(m)}(t)\; b^{(k)}(t)\;, \end{aligned} \tag{23}$$

where the number of modes used in the velocity and temperature decompositions are $N_u = 20$ and $N_T = 6$, respectively. As discussed previously, these numbers of modes ensure that 90% of the total variance is captured in both decompositions. The superscript $\dot{}$ denotes a time derivative. The $\hat{}$ notation is introduced here to distinguish these dynamical representations from the temporal modes calculated from the original data. The temporal velocity POD modes, $\hat{a}^{(n)}(t)$, are dynamically evolved, whilst the temperature modes, $b^{(n)}(t)$, are prescribed akin to a source term. The superscripts on the terms D_n^{ϑ}, C_n^{ϑ}, $F_n^{\vartheta\vartheta'}$, L_{nm}^{ϑ}, and Z_{nm}^{ϑ}, denotes that these coefficients are symbolically functions of seasonally dependent fields, as expanded upon in the following paragraph.

The coefficients in (23) can all in principle be calculated from the spatial POD modes and phase averaged fields, using the expressions provided below. Note, we do not do so here, but these symbolic representations provide the justifications for the input factors used in the regression approach to calculating these coefficients. Firstly, D_n^{ϑ} is associated with the time derivative of the seasonal cycle being brought over from the left hand side of (22), and is given by

$$D_n^{\vartheta} \;=\; \left\langle \boldsymbol{\mathcal{U}}^{(n)}, -\widetilde{\mathbf{u}}^{\vartheta} \right\rangle_u \quad . \tag{24}$$

The coefficients involving linear and quadratic functions of the seasonal cycle, are respectively given by

$$C_n^{\vartheta} \;=\; \left\langle \boldsymbol{\mathcal{U}}^{(n)}, \boldsymbol{\mathcal{L}}(\breve{\mathbf{u}}^{\vartheta}) + \boldsymbol{\mathcal{S}}(\breve{T}^{\vartheta}) \right\rangle_u \quad \text{, and} \tag{25}$$

$$F_n^{\vartheta\vartheta'} \;=\; \left\langle \boldsymbol{\mathcal{U}}^{(n)}, \boldsymbol{\mathcal{N}}(\breve{\mathbf{u}}^{\vartheta}, \breve{\mathbf{u}}^{\vartheta'}) + \boldsymbol{\mathcal{M}}(\breve{T}^{\vartheta}, \breve{T}^{\vartheta'}) \right\rangle_u \quad . \tag{26}$$

Note, due to the properties of $d^{\vartheta}(t)$, on the raw data $d^{\vartheta}(t)d^{\vartheta'}(t) = \delta_{\vartheta\vartheta'}$ for all t. This means that $F_n^{\vartheta\vartheta'}$ is only non-zero when $\vartheta = \vartheta'$. However, when applied to the temporally interpolated time series, $d^{\vartheta}(t)d^{\vartheta'}(t)$, and hence $F_n^{\vartheta\vartheta'}$, are strictly speaking also non-zero for adjacent months such that $|\vartheta - \vartheta'| \leq 1$. The linear velocity POD

terms are

$$L^{\vartheta}_{nm} = \left\langle \boldsymbol{\mathcal{U}}^{(n)}, \boldsymbol{\mathcal{N}}(\boldsymbol{\mathcal{U}}^{(m)}, \breve{\mathbf{u}}^{\vartheta}) + \boldsymbol{\mathcal{N}}(\breve{\mathbf{u}}^{\vartheta}, \boldsymbol{\mathcal{U}}^{(m)}) + \boldsymbol{\mathcal{L}}(\boldsymbol{\mathcal{U}}^{(m)}) \right\rangle_u \quad . \tag{27}$$

The velocity POD quadratic terms have no seasonal dependence justified by the equations of motion and are given by

$$Q_{nmk} = \left\langle \boldsymbol{\mathcal{U}}^{(n)}, \boldsymbol{\mathcal{N}}(\boldsymbol{\mathcal{U}}^{(m)}, \boldsymbol{\mathcal{U}}^{(k)}) \right\rangle_u \quad . \tag{28}$$

Note, by virtue of the symmetry properties in summation of these quadratic terms in the (23), one can redefine Q_{nmk} to be $(Q_{nmk} + Q_{nkm})/2$, without loss of generality. The linear temperature POD mode coefficients are

$$Z^{\vartheta}_{nm} = \left\langle \boldsymbol{\mathcal{U}}^{(n)}, \boldsymbol{\mathcal{M}}(\mathcal{T}^{(m)}, \breve{T}^{\vartheta}) + \boldsymbol{\mathcal{M}}(\breve{T}^{\vartheta}, \mathcal{T}^{(m)}) + \boldsymbol{\mathcal{S}}(\mathcal{T}^{(m)}) \right\rangle_u \quad . \tag{29}$$

Finally the temperature POD quadratic terms also have no seasonal dependence and are given by

$$R_{nmk} = \left\langle \boldsymbol{\mathcal{U}}^{(n)}, \boldsymbol{\mathcal{M}}(\mathcal{T}^{(m)}, \mathcal{T}^{(k)}) \right\rangle_u \quad . \tag{30}$$

Again, one can redefine R_{nmk} to be $(R_{nmk} + R_{nkm})/2$, without loss of generality. To determine the POD ROM coefficients using the above equations, one must calculate sufficiently accurate spatial derivatives. This can be particularly difficult for complex geometries on arbitrary grids. Here we adopt an alternate approach, where these coefficients are instead calculated using linear regression, as outlined in the following section.

Given the above symbolic representations, one can define a set of physically meaningful subsets of coefficients. Each subset is associated with simplified versions of the equation of state linking density to temperature. To represent the full complexity of the quadratic equation of state all of the coefficients are required, namely C^{ϑ}_n, $F^{\vartheta\vartheta'}_n$, D^{ϑ}_n, L^{ϑ}_{nm}, Q_{nmk}, the seasonally varying linear temperature coefficients Z^{ϑ}_{nm}, and season invariant quadratic temperature coefficients R_{nmk}. We also test variants when quadratic temperature coefficients R_{nmk} are excluded, to determine their relative importance. One can also define an equation of state in which density is only linearly dependent upon temperature by excluding the $\boldsymbol{\mathcal{M}}$ operator. In this instance the only non-zero coefficients are C^{ϑ}_n, $F^{\vartheta\vartheta'}_n$, D^{ϑ}_n, L^{ϑ}_{nm}, Q_{nmk} and a season invariant linear temperature coefficient Z_{nm}. Finally when all temperature terms are removed, the only non-zero coefficients are C^{ϑ}_n, $F^{\vartheta\vartheta'}_n$, D^{ϑ}_n, L^{ϑ}_{nm} and Q_{nmk}. In this case density can only be depth dependent. We also additionally test all of the above variants without the nonlinear meanfield coefficients $F^{\vartheta\vartheta'}_n$, and without both coefficients D^{ϑ}_n and $F^{\vartheta\vartheta'}_n$.

5 Calculation of the POD ROM coefficients

The true value of the POD temporal modes and their time derivatives are known from the raw data. Based on the form of the ODEs in (23), one can then form the following

regression problem

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} \ , \tag{31}$$

solved separately for each mode n. The matrix $\mathbf{Y}$ contains the time derivatives, $\dot{a}^{(n)}(t)$, as per the left-hand-side of (23). When solving for the full complexity, $\boldsymbol{\beta}$ contains the POD ROM coefficients (D_n^ϑ, C_n^ϑ, $F_n^{\vartheta\vartheta'}$, L_{nm}^ϑ, Q_{nmk}, Z_{nm}^ϑ, R_{nmk}), including only the unique coefficients spanning all m and k. $\mathbf{X}$ then contains the terms multiplied by these coefficients in the right-hand-side of (23). The structure of $\mathbf{Y}$, $\mathbf{X}$ and $\boldsymbol{\beta}$ are detailed in appendix B. The time series used to construct $\mathbf{X}$ are also scaled such that they both have zero mean and are approximately bounded from -1 to 1. This scaling is important to ensure the regularisation discussed below appropriately balances between the fit to data, and size of the parameters. Solving for $\boldsymbol{\beta}$ in (31) yields the required POD ROM coefficients.

Here we solve for $\boldsymbol{\beta}$ using ridge regression, where L2 regularization is applied to the squared magnitude of the POD ROM coefficients. The cost function to be minimised with respect to $\boldsymbol{\beta}$, is defined as

$$\mathcal{J}(\boldsymbol{\beta}) = \frac{1}{N}\left(\mathbf{Y} - \boldsymbol{\beta}\mathbf{X}\right)\left(\mathbf{Y} - \boldsymbol{\beta}\mathbf{X}\right)^T + \kappa\boldsymbol{\beta}\boldsymbol{\beta}^T \ , \tag{32}$$

where N is the number of samples used over the training period, and κ is the ridge regression hyper-parameter. This cost function has the closed form solution

$$\boldsymbol{\beta} = \frac{\mathbf{Y}\mathbf{X}^T}{N}\left(\frac{\mathbf{X}\mathbf{X}^T}{N} + \kappa\mathbf{I}\right)^{-1} \ , \tag{33}$$

where the superscript ^{T} denotes the transpose operation. $\mathbf{I}$ is the identity matrix of leading dimension equal to the number of input factors. We calculate (33) using the monthly averaged fields across the entire CAFE-60 ensemble from 2010 to 2015. The data is temporally interpolated to a time step of $1/30^{\text{th}}$ of a month, or approximately one day. This finer time step size is required for the numerical stability of the POD ROM in the following section. Cubic splines are used to interpolate $a^{(n)}(t)$, which are analytically differentiated to determine compatible values for $\dot{a}^{(n)}(t)$. Cubic splines are also used to interpolate $b^{(n)}(t)$. The $d^\vartheta(t)$ time series are interpolated such that they transition from 0 in the previous month to 1 in the current month to 0 again in the following month using a scaled and offset cosine function. This is analytically differentiated to produce interpolated versions of $\dot{d}^\vartheta(t)$. After this temporal interpolation, there are $172,800$ samples or data instances used to solve the regression problem (5 years $\times$ 12 months per year $\times$ 30 time steps per month $\times$ 96 ensemble members).

The following error statistics comparing the predicted and actual values of $\mathbf{Y}$ are calculated, per mode n, and per hyper-parameter κ. These statistics are calculated over the out-of-sample period from 2015 to 2020, using the first data instance of each month (void of any temporal interpolation). Since the samples across the ensemble are not necessarily independent, we compare the predicted and actual values on the basis of their average across the ensemble.

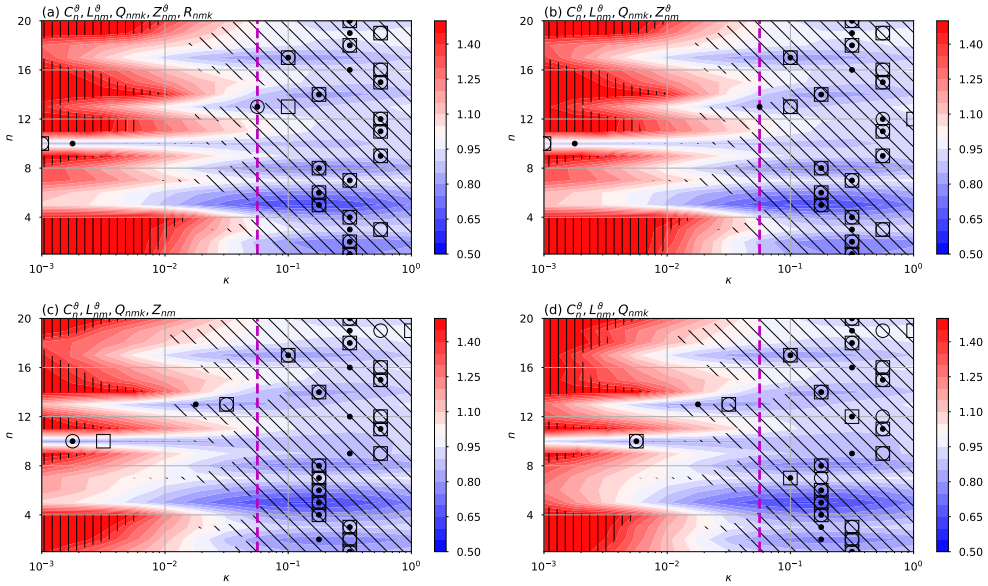

Fig. 5 Root mean squared error of ridge regression normalised by the variance of the output per mode n for various hyper-parameters κ when solving for coefficients: (a) C_n^ϑ, L_{nm}^ϑ, Q_{nmk} excluding any temperature dependence; (b) C_n^ϑ, L_{nm}^ϑ, Q_{nmk} with fixed in time linear temperature coefficients Z_{nm}; (c) C_n^ϑ, L_{nm}^ϑ, Q_{nmk} with climatological linear time temperature coefficients Z_{nm}^ϑ; and (d) C_n^ϑ, L_{nm}^ϑ, Q_{nmk} with climatological linear time temperature coefficients Z_{nm}^ϑ, and fixed in time quadratic temperature coefficients R_{nmk}. Solid black line is where the mean squared error is the same as the variance in the predicted output. Filled symbols locate the hyper-parameter of minimum MSE per mode. Hollow circle symbols locate the hyper-parameter of minimum MSE per mode, when additionally solving for D_n^ϑ. Hollow square symbols locate the hyper-parameter of minimum MSE per mode, when additionally solving for D_n^ϑ and $F_n^{\vartheta\vartheta'}$. $\kappa = 0.056$ is indicated by the vertical magenta line.

The ability to represent $\mathbf{Y}$ is tested using all of the theoretically justified parameters, and also the subsets defined in the previous section. Each subset is associated to simplified versions of the equation of state. We initially solve the regression problem with the exclusion of both the seasonal time derivative term D_n^ϑ, and the nonlinear meanfield term $F_n^{\vartheta\vartheta'}$. Figure 5 illustrates the root mean squared error (RMSE) of mode n per hyper-parameter κ. We assess κ over a logarithmically spaced grid from 1×10^{-3} to 1. The RMSE is normalised by the standard deviation of the actual time derivative of mode n over the test period. Figure 5(a) presents the error measures when solving for the full complement of remaining coefficients associated with an equation of state quadratic in temperature. In figure 5(b) we test the model when quadratic temperature coefficients R_{nmk} are excluded. Figure 5(c) solves for C_n^ϑ, L_{nm}^ϑ, Q_{nmk} and a fixed in time linear temperature coefficient Z_{nm}, associated with a linear dependence of density on temperature. Figure 5(d) presents the results when we are only solving for the parameters C_n^ϑ, L_{nm}^ϑ and Q_{nmk}, associated with a density field with no temperature dependence.

The red shading in these figures are where the RMSE is greater than the standard deviation, and hence have worse skill than continually predicting the mean. The vertical hatching are regions in which the RMSE is greater than the standard deviation to a 99% confidence level. The blue shading indicate where the RMSE is less than the standard deviation, and are potentially skilful. The black dots in these figures indicate the hyper-parameter of minimum RMSE per mode, and are all located within the blue regions. Each set of calculations are repeated by solving for the sets of parameters additionally including D_n^ϑ. The hollow circle symbols in each plot indicate the associated hyper-parameter of minimum RMSE. Likewise, the calculations are repeated where both D_n^ϑ and $F_n^{\vartheta\vartheta'}$ are additionally included, with the hollow square symbols locating the hyper-parameter of minimum RMSE. The optimal hyper-parameter is not changed in many instances with the inclusion of these additional terms. Those instances where it has changed, it has moved by only one evaluated value of κ.

The blue regions have higher hyper-parameters than the red regions, and hence penalise the coefficient magnitude to a greater extent. Whilst the parameter sets in the blue zone have a lower RMSE, they also produce lower variability than that of the actual time derivatives. The diagonal hatching indicates regions where the variance of the predicted time derivative is less than that of the actual to a 99% confidence level. This is not a desired property if one is looking to ensure the variance of the predicted POD ROM per mode is sufficiently similar to that of the underlying data. With the exception of mode $n = 10$, the optimal hyper-parameters all lie in regions of insufficient variance. There is in fact only a narrow lane of hyper-parameters that lie outside of both zones of insufficient agreement (vertical hatch), and insufficient variance (diagonal hatch). Since we solve for each POD mode individually, one could in principle have a different hyper-parameter for each mode n. For the sake of simplicity, however, we select $\kappa = 0.056$ for all modes and all model variants, as it is the lowest value inside the blue zone for the majority of modes for each variant. This value of κ is indicated by the vertical magenta line in figure 5.

6 POD ROM temporal integration

The RMSE statistics presented in the previous section are effectively how well the set of coefficients represents the time derivative of the temporal POD mode one time step into the future. The tougher test, and ultimate goal, is how well one can reproduce the dynamically evolving temporal POD modes simulated throughout the observed period. The coefficients are learnt across all of the samples using $\kappa = 0.056$, and for each of the aforementioned subsets of parameters. We adopt these POD ROM coefficients in (23), which is solved using the fourth order Runge-Kutta time stepping scheme. We tested for the influence of temporal discretisation error, by running experiments across a variety of time step sizes. The results presented in this manuscript are insensitive to the choice of time steps smaller than the adopted one of $1/30^{\text{th}}$ of a month.

Figure 6 illustrates the temporal integrations of the first 6 POD modes. POD ROMs with equations of state that have either no temperature dependence (C_n^ϑ, L_{nm}^ϑ, Q_{nmk} - red lines), or only a linear one (C_n^ϑ, L_{nm}^ϑ, Q_{nmk}, Z_{nm} - green lines) have a poorer fit to the original data. The variants that retain some elements of the quadratic

relationship between density and temperature (blue and magenta lines), are in better agreement across all of the modes. For all variants the solid line represent a POD ROM with no seasonal derivative terms (D_n^ϑ) nor any nonlinear meanfield terms ($F_n^{\vartheta\vartheta'}$). The dotted lines include D_n^ϑ, and the dashed lines include both D_n^ϑ and $F_n^{\vartheta\vartheta'}$. The addition of the seasonal derivative and nonlinear meanfield terms have a negligible influence on the evolution of all systems. Recall our POD ROM represents the evolution of the perturbations about the seasonal cycle. The fact that D_n^ϑ and $F_n^{\vartheta\vartheta'}$ provided negligible improvement, suggests that these perturbations are less influenced by the seasonal derivatives and nonlinear seasonal interactions, and more so by the linear and nonlinear interactions between the perturbations themselves.

These observations are also evident in the statistical comparisons between the underlying data and the POD ROM integrations per mode. Figure 7(a) illustrates a comparison of the POD kinetic energy spectra (black line), and each of the various POD ROMs. For the moment we concentrate on the solid lines, which are the POD ROM variants not including the D_n^ϑ and $F_n^{\vartheta\vartheta'}$ terms. The variants retaining some quadratic relationship between density and temperature (blue and magenta) are able to reproduce the energy across the POD spectrum. These variants also have correlations between the integrated and actual time series of above 0.8 for the first 11 modes, as indicated in figure 7(b). Their root-mean-squared-error (RMSE) is also lower than the remaining variants, as shown in figure 7(c). The prescribed temperature POD modes, $b^{(n)}(t)$, are key to the agreement between the temporal integrations $\hat{a}^{(n)}(t)$, and the original POD modes $a^{(n)}(t)$. The anomalous temperature field is setting the phase of the large scale structures, to which the anomalous velocity field responds. As such, it is unsurprising that the variants with no influence from the temperature field (red lines) have the lowest correlation and highest RMSE.

The inclusion of the D_n^ϑ and $F_n^{\vartheta\vartheta'}$ terms have made little discernible impact on the energy per mode in figure 7(a) for all sets of parameters. Figure 7(b) illustrates that the inclusions of these terms has resulted in a small decrease in correlation across all parameter sets for most modes. Consistently, figure 7(c) indicates that the normalised RMSE has increased in most instances with the inclusion of these additional parameters. This suggests that whilst these terms are theoretically justified, they have negligible impact on the physical system. The inclusion of D_n^ϑ and $F_n^{\vartheta\vartheta'}$ in the regression problem, does not add to the performance of the POD ROM, but rather exacerbates the problem of learning the other significant coefficients from the available samples.

7 Concluding remarks

To summarise, a POD ROM was developed for the global oceans on monthly averaged time scales over a recent decade. As opposed to black-box methods, the data-driven approach adopted here is both physically constrained and interpretable. Three-dimensional POD modes were calculated from the ensemble averaged ocean fields of the CAFE-60 reanalysis. The contribution of these modes to the 96 individual ensemble members of CAFE-60 was determined. A reduced-order dynamical system was constructed to be constrained by the physics, by projecting the hydrostatic Boussinesq

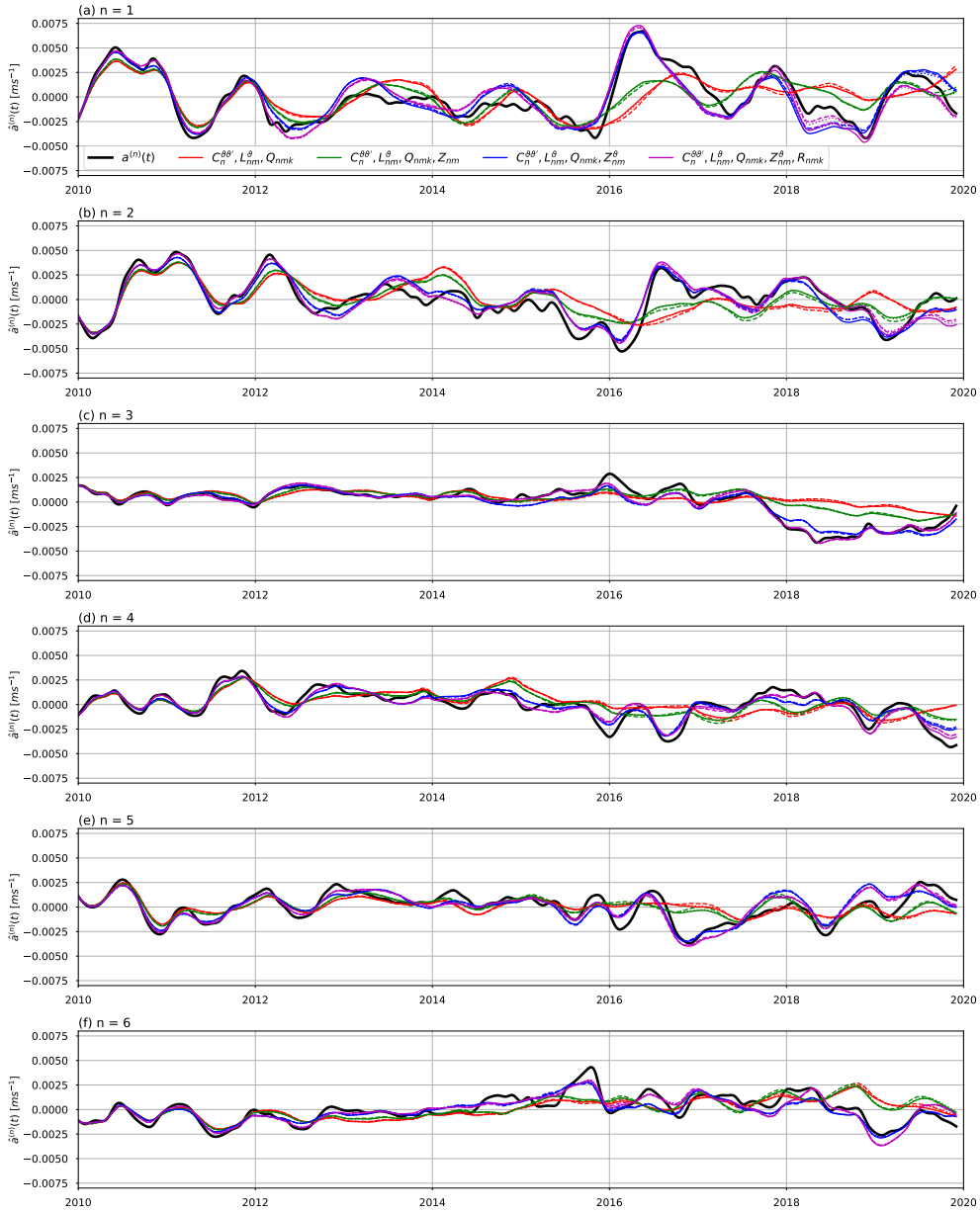

Fig. 6 POD ROM temporal integration of the ensemble average adopting subsets of model coefficients for modes: (a) $n = 1$; (b) $n = 2$; (c) $n = 3$; (d) $n = 4$; (e) $n = 5$; and (f) $n = 6$. The legend in (a) associating the subsets of model coefficients is applicable to all figures. Dotted lines additionally include the D_n^{ϑ} terms. Dashed lines additionally include the D_n^{ϑ} and $F_n^{\vartheta\vartheta'}$ terms.

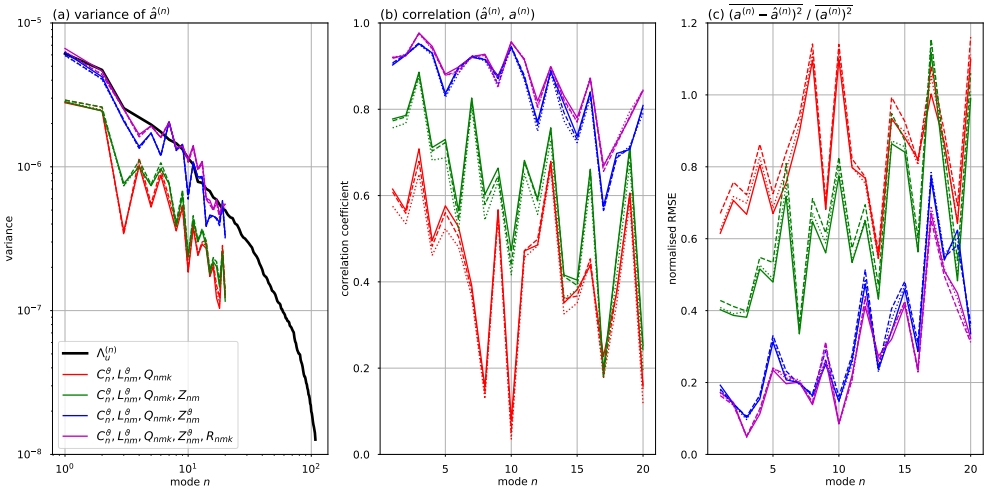

Fig. 7 Comparative statistics per mode of the POD ROMs on the basis of: (a) variance; (b) correlation between $\hat{a}^{(n)}(t)$ and $a^{(n)}(t)$; and (c) normalised root-mean-square-error. The legend in (a) associating the subsets of model coefficients is applicable to all figures. Dotted lines additionally include the D_n^ϑ terms. Dashed lines additionally include the D_n^ϑ and $F_n^{\vartheta\vartheta'}$ terms.

equations of motion onto a horizontal velocity POD mode basis. A set of temperature POD modes acted as a prescribed source term. The coefficients in this POD ROM were determined via a regression approach, exploiting all of the samples in time and across the full 96 member ensemble. Various POD ROMs were tested each with effectively different equations of state. Each variant was evaluated by its ability to simulate the ensemble averaged field. POD ROM variants with a quadratic dependence between the density and temperature fields were able to reproduce the statistics of the large scale structures. This was achieved at a mere fraction of the computational cost required by a GCM to simulate such a flow. The inclusion of input factors associated with the seasonal cycle time derivative and nonlinear meanfield terms, was found to not improve performance.

The seasonal component was represented here as statistics per calendar month, which is the standard approach in the climate sciences. This was done to enable the results and model coefficients to be interpreted in this standard way. However, the calendar months of the year are human constructs. Alternatively, an approach potentially more faithful to the physics, would be to create "phase angles" that are instead centred around the solstices and equinoxes linked to the orbit of the Earth. We will investigate this approach in future studies.

Finally, having estimates of the evolution of the three-dimensional ocean velocity field has numerous real-world applications. For instance it may enable forecasting the transport of various scalar quantities, including: biological species; chemical spills; and plastic pollution.

Declarations

The authors acknowledge funding from the CSIRO Artificial Intelligence for Missions program. CAFE-60 ensemble reanalysis dataset is publicly and freely available at https://registry.opendata.aws/csiro-cafe60/.

Appendix A Hydrostatic Boussinesq equations of motion

The hydrostatic Boussinesq equations of motion constitute the conservation of horizontal momentum, hydrostatic balance, conservation of mass, and an equation of state [37]. In the following notation, time is denoted by t. The spatial coordinates $\mathbf{x} \equiv (x, y, z)$ are attached to the rotating Earth, and positive pointing east, north and up, respectively. The Earth's surface being located at $z = 0$. The zonal (south to north), meridional (west to east) and vertical velocity components are denoted by u, v and w.

The Boussinesq approximation decomposes the density (ρ) and pressure (p) fields according to

$$\rho(x, y, z, t) = \check{\rho}(x, y, z, t) + \mathring{\rho}(z) + \rho_0 \text{ , and} \tag{A1}$$
$$p(x, y, z, t) = \check{p}(x, y, z, t) + \mathring{p}(z) \text{ ,} \tag{A2}$$

where one assumes that $\rho_0 \gg |\mathring{\rho}|, |\check{\rho}|$. This is valid in the ocean, with $\rho_0 \approx 1.04$ g cm^{-3}, $|\mathring{\rho}| \approx 0.03$ g cm^{-3} and $|\check{\rho}| \approx 0.003$ g cm^{-3} [37]. The density field is approximated to ρ_0 in all evolution equations, other than for the gravitational (or buoyancy) term.

Under these assumptions the conservation of horizontal momentum equations are

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + w\frac{\partial u}{\partial z} = -\frac{1}{\rho_0}\frac{\partial \check{p}}{\partial x} + fv \text{ , and} \tag{A3}$$
$$\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + w\frac{\partial v}{\partial z} = -\frac{1}{\rho_0}\frac{\partial \check{p}}{\partial y} - fu \text{ ,} \tag{A4}$$

where the Coriolis parameter $f = 2\Omega \sin\phi$, of latitude ϕ, and $\Omega = 7.292 \times 10^{-5}$s^{-1} the angular velocity magnitude of the Earth. Note, the viscous term is negligible for the global ocean at the scales resolved in the current reanalysis.

In the global ocean the vertical velocity is orders of magnitude smaller than the other velocity components. Applying this to the vertical momentum equation, results in the hydrostatic approximation

$$\frac{\partial p}{\partial z} = -\rho g \text{ .} \tag{A5}$$

Since there is now no time derivative term in the w velocity component momentum equation, w is a diagnostic field as opposed to a prognostic one. The time invariant components of pressure and density satisfy the hydrostatic approximation in (A5) such

that

$$\frac{\partial \mathring{p}(z)}{\partial z} = -\left(\rho_0 + \mathring{\rho}(z)\right) g \quad . \tag{A6}$$

Subtracting (A6) away from (A5) produces the hydrostatic equation for the perturbation components given by

$$\frac{\partial \check{p}(x,y,z,t)}{\partial z} = -\check{\rho}(x,y,z,t) g \quad . \tag{A7}$$

Integrating both sides of (A7) produces

$$\check{p}(x,y,z,t) = -g \int_z^0 \check{\rho}(x,y,\underline{z},t) d\underline{z} \quad , \tag{A8}$$

giving an expression for the perturbation component of pressure. Here $\underline{z}$ is used to distinguish the terms in the integrand from the lower limit in the integral. Note, z is the lower limit because the ocean surface is at $z = 0$, and since z is positive upward, the ocean depths have negative values of z.

The conservation of mass is given by

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \quad , \tag{A9}$$

which can be rearranged as follows

$$w(x,y,z,t) = -\int_z^0 \left(\frac{\partial u(x,y,\underline{z},t)}{\partial x} + \frac{\partial v(x,y,\underline{z},t)}{\partial y} \right) d\underline{z} \ , \tag{A10}$$

to obtain a diagnostic equation for the vertical velocity.

There are no fundamental equations of state for seawater, but rather a family of empirical ones. In general the density of seawater is a nonlinear relationship involving temperature, pressure and salinity, with the latter playing a lesser role [38]. Note, we do not require an equation of state for the complete density field, ρ, instead only one for the perturbation component, $\check{\rho}$. This requirement effectively removes the direct influence of pressure. We then adopt a modified version of [39] excluding salinity effects such that

$$\check{\rho}(x,y,z,t) \ = \ -\gamma_1 z T(x,y,z,t) - \gamma_2 (T(x,y,z,t) - T_{\rm ref})^2 \ . \tag{A11}$$

This retains the minimal complexity in the relationship with temperature, where γ_1, γ_2 and $T_{\rm ref}$ are empirically derived parameters. Substituting (A11) into (A8) produces

$$\check{p}(x,y,z,t) = g \int_z^0 \left[\gamma_1 \underline{z} T(x,y,\underline{z},t) + \gamma_2 (T(x,y,\underline{z},t) - T_{\rm ref})^2 \right] d\underline{z} \quad , \tag{A12}$$

which is now an expression for the perturbation component of pressure as a function of temperature. Note, γ_1, γ_2 and T_{ref} can also be depth dependent, and not influence the following derivation.

Substituting (A12) for $\check{p}$, and (A10) for w, into the horizontal momentum equations (A3) and (A4), produces

$$\begin{aligned}\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y} &- \int_z^0\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}\right)d\underline{z}\frac{\partial u}{\partial z}=fv\\ &-\frac{g}{\rho_0}\int_z^0\left[\gamma_1\underline{z}+2\gamma_2(T-T_{\text{ref}})\right]\frac{\partial T}{\partial x}d\underline{z}\ , \text{ and} \qquad \text{(A13)}\\ \frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y} &- \int_z^0\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}\right)d\underline{z}\frac{\partial v}{\partial z}=-fu\\ &-\frac{g}{\rho_0}\int_z^0\left[\gamma_1\underline{z}+2\gamma_2(T-T_{\text{ref}})\right]\frac{\partial T}{\partial y}d\underline{z}\ . \qquad \text{(A14)}\end{aligned}$$

The above two equations have three unknowns, u, v and T. At this stage we produce a POD ROM for the velocity field, with the temperature field entering as a prescribed source term. Future work will also have a dynamically evolving temperature field, which would additionally require the adoption of the energy equation.

Making use of the unit vectors $\mathbf{e}_1=(1,0)$ and $\mathbf{e}_2=(0,1)$ we can write the horizontal momentum equation in the vector form

$$\begin{aligned}\frac{\partial\mathbf{u}}{\partial t} &= \boldsymbol{\mathcal{N}}(\mathbf{u},\mathbf{u})+\boldsymbol{\mathcal{L}}(\mathbf{u})+\boldsymbol{\mathcal{M}}(T,T)+\boldsymbol{\mathcal{S}}(T)\ , \text{ where} \qquad \text{(A15)}\\ \boldsymbol{\mathcal{N}}(\mathbf{u},\mathbf{v}) &= -\mathbf{u}\cdot\boldsymbol{\nabla}_h\mathbf{v}+\int_z^0\boldsymbol{\nabla}_h\cdot\mathbf{u}\ d\underline{z}\ \frac{\partial\mathbf{v}}{\partial z}\ , \qquad \text{(A16)}\\ \boldsymbol{\mathcal{L}}(\mathbf{u}) &= f\left[(\mathbf{e}_2\cdot\mathbf{u})\mathbf{e}_1-(\mathbf{e}_1\cdot\mathbf{u})\mathbf{e}_2\right]\ , \qquad \text{(A17)}\\ \boldsymbol{\mathcal{M}}(T,\Upsilon) &= -\frac{g}{\rho_0}\int_z^0 2\gamma_2 T\boldsymbol{\nabla}_h\Upsilon d\underline{z}\ , \qquad \text{(A18)}\\ \boldsymbol{\mathcal{S}}(T) &= -\frac{g}{\rho_0}\int_z^0(\gamma_1\underline{z}-2\gamma_2 T_{\text{ref}})\boldsymbol{\nabla}_h T d\underline{z}\ , \text{ and} \qquad \text{(A19)}\\ \boldsymbol{\nabla}_h &= \left(\frac{\partial}{\partial x},\frac{\partial}{\partial y}\right)\ , \qquad \text{(A20)}\end{aligned}$$

is the horizontal derivative operator, with $\mathbf{u}=(u,v)$. The quantities $\mathbf{v}$ and ~~r~~$\underline{\Upsilon}$, are of the same type as $\mathbf{u}$ and T, respectively, and have been introduced here for clarity of notation. Note, if the viscous term was included, it would appear in the expression for $\boldsymbol{\mathcal{L}}(\mathbf{u})$.

Appendix B Calculation of POD ROM coefficients using regression

In the regression approach to calculating the POD ROM coefficients, here we solve for $\boldsymbol{\beta}$ in the equation

$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{X} \quad , \tag{B21}$$

for each POD mode n, separately. In the most general application, involving the calculation of all parameters in the physically justified POD ROM

$$\mathbf{Y} = \dot{\mathbf{a}}^{(n)} \quad \text{, and} \tag{B22}$$

$$\boldsymbol{\beta} = \left[\mathbf{D}_n \middle| \mathbf{C}_n \middle| \mathbf{F}_n \middle| \mathbf{L}_n^{\text{Jan}} \cdots \mathbf{L}_n^{\vartheta} \cdots \mathbf{L}_n^{\text{Dec}} \middle| \mathbf{Q}_n \middle| \mathbf{Z}_n^{\text{Jan}} \cdots \mathbf{Z}_n^{\vartheta} \cdots \mathbf{Z}_n^{\text{Dec}} \middle| \mathbf{R}_n \right] , \tag{B23}$$

where the coefficient vectors in $\boldsymbol{\beta}$ are defined as

$$\mathbf{D}_n = \left[D_n^{\text{Jan}} \cdots D_n^{\vartheta} \cdots D_n^{\text{Dec}} \right] , \tag{B24}$$

$$\mathbf{C}_n = \left[C_n^{\text{Jan}} \cdots C_n^{\vartheta} \cdots C_n^{\text{Dec}} \right] , \tag{B25}$$

$$\mathbf{F}_n = \left[F_n^{\text{Jan Jan}} \cdots F_n^{\vartheta\vartheta'} \cdots F_n^{\text{Dec Dec}} \right] , \tag{B26}$$

$$\mathbf{L}_n^{\vartheta} = \left[L_{n1}^{\vartheta} \cdots L_{nm}^{\vartheta} \cdots L_{nN_u}^{\vartheta} \right] , \tag{B27}$$

$$\mathbf{Q}_n = \left[Q_{n11} \cdots Q_{nmk} \cdots Q_{nN_uN_u} \right] , \tag{B28}$$

$$\mathbf{Z}_n^{\vartheta} = \left[Z_{n1}^{\vartheta} \cdots Z_{nm}^{\vartheta} \cdots Z_{nN_T}^{\vartheta} \right] \text{, and} \tag{B29}$$

$$\mathbf{R}_n = \left[R_{n11} \cdots R_{nmk} \cdots R_{nN_TN_T} \right] \quad . \tag{B30}$$

The $F_n^{\vartheta\vartheta'}$ coefficients are only included for adjacent months such that $|\vartheta - \vartheta'| \leq 1$. The vector $\mathbf{Q}_n$ only include the unique coefficients of Q_{nmk}, recalling that $Q_{nmk} = Q_{nkm}$. Likewise $\mathbf{R}_n$ only includes the unique coefficients of R_{nmk}, where $R_{nmk} = R_{nkm}$. The vertical lines in (B23), separate $\boldsymbol{\beta}$ into a set of seven blocks. The associated seven

input factors blocks are separated by horizontal lines in the definition of $\mathbf{X}$, given by

$$\mathbf{X} = \begin{bmatrix} \dot{\mathbf{d}}^{\text{Jan}} \\ \vdots \\ \dot{\mathbf{d}}^{\vartheta} \\ \vdots \\ \dot{\mathbf{d}}^{\text{Dec}} \\ \hline \mathbf{d}^{\text{Jan}} \\ \vdots \\ \mathbf{d}^{\vartheta} \\ \vdots \\ \mathbf{d}^{\text{Dec}} \\ \hline \mathbf{d}^{\text{Jan}} \odot \mathbf{d}^{\text{Jan}} \\ \vdots \\ \mathbf{d}^{\vartheta} \odot \mathbf{d}^{\vartheta'} \\ \vdots \\ \mathbf{d}^{\text{Dec}} \odot \mathbf{d}^{\text{Dec}} \\ \hline \mathbf{A}^{\text{Jan}} \\ \vdots \\ \mathbf{A}^{\vartheta} \\ \vdots \\ \mathbf{A}^{\text{Dec}} \\ \hline \mathbf{a}^{(1)} \odot \mathbf{a}^{(1)} \\ \vdots \\ \mathbf{a}^{(m)} \odot \mathbf{a}^{(k)} \\ \vdots \\ \mathbf{a}^{(N_u)} \odot \mathbf{a}^{(N_u)} \\ \hline \mathbf{B}^{\text{Jan}} \\ \vdots \\ \mathbf{B}^{\vartheta} \\ \vdots \\ \mathbf{B}^{\text{Dec}} \\ \hline \mathbf{b}^{(1)} \odot \mathbf{b}^{(1)} \\ \vdots \\ \mathbf{b}^{(m)} \odot \mathbf{b}^{(k)} \\ \vdots \\ \mathbf{b}^{(N_T)} \odot \mathbf{b}^{(N_T)} \end{bmatrix}, \text{ where} \tag{B31}$$

with

$$\mathbf{A}^{\vartheta} = \begin{bmatrix} \mathbf{a}^{(1)} \odot \mathbf{d}^{\vartheta} \\ \vdots \\ \mathbf{a}^{(m)} \odot \mathbf{d}^{\vartheta} \\ \vdots \\ \mathbf{a}^{(N_u)} \odot \mathbf{d}^{\vartheta} \end{bmatrix} \text{, and} \tag{B32}$$

$$\mathbf{B}^{\vartheta} = \begin{bmatrix} \mathbf{b}^{(1)} \odot \mathbf{d}^{\vartheta} \\ \vdots \\ \mathbf{b}^{(m)} \odot \mathbf{d}^{\vartheta} \\ \vdots \\ \mathbf{b}^{(N_T)} \odot \mathbf{d}^{\vartheta} \end{bmatrix} , \tag{B33}$$

with $\odot$ denoting the Hadamard (or Schur) product for component-wise multiplication. The horizontal vectors $\mathbf{a}^{(n)}$, $\mathbf{b}^{(n)}$, $\mathbf{d}^{\vartheta}$, $\dot{\mathbf{a}}^{(n)}$, and $\dot{\mathbf{d}}^{\vartheta}$ contain the interpolated and transformed versions of the temporal elements of $a^{(n)}(t)$, $b^{(n)}(t)$, $d^{\vartheta}(t)$, $\dot{a}^{(n)}(t)$ and $\dot{d}^{\vartheta}(t)$, respectively. The products $\mathbf{d}^{\vartheta} \odot \mathbf{d}^{\vartheta'}$ in $\mathbf{X}$ are associated with the $F_n^{\vartheta\vartheta'}$ coefficients, and hence are only included for adjacent months where $|\vartheta - \vartheta'| \leq 1$. The product terms $\mathbf{a}^{(m)} \odot \mathbf{a}^{(k)}$ and $\mathbf{b}^{(m)} \odot \mathbf{b}^{(k)}$, are associated with Q_{nmk} and R_{nmk}, respectively, and only include unique combinations of these products.

For increased temporal resolution cubic splines are used to interpolate $a^{(n)}(t)$ and also determine compatible values for $\dot{a}^{(n)}(t)$. Cubic splines are also used to interpolate $b^{(n)}(t)$. The $d^{\vartheta}(t)$ time series are interpolated such that they transition from 0 in the previous month to 1 in the current month to 0 again in the following month using a scaled and offset cosine function. This is analytically differentiated to produce interpolated versions of $\dot{d}^{\vartheta}(t)$. All of these time series are transformed to have approximately zero mean and bounded from -1 to 1. The rows of $\mathbf{Y}$ and $\mathbf{X}$ are horizontally concatenated to include the samples across the 96 member ensemble. This only increases the number of samples, and does not change the dimensions of $\boldsymbol{\beta}$.

In principle one could use any appropriate methods to solve for $\boldsymbol{\beta}$. Here we adopt ridge regression. When solving for a subset of the parameters, one removes the parameters that are not required from the $\boldsymbol{\beta}$, and also remove the associated input factors from $\mathbf{X}$.

References

[1] Holton, J.R.: An Introduction to Dynamic Meteorology. Elsevier, New York (2004)

[2] Kitsios, V., O'Kane, T.J., Zagar, N.: A reduced order representation of the madden-julian oscillation based on reanalyzed normal mode coherences. J. Atmos. Sci. **76**, 2463–2480 (2019)

[3] O'Kane, T.J., Kitsios, V., Collier, M.A.: On the semiannual formation of large scale three-dimensional vortices at the stratopause. Geophys. Res. Let. **48**(4), 2020–090072 (2020)

[4] Einstein, A.: "How can we save mankind and it's spiritual acquisitions of which we are the heirs and how can one save Europe from a new disaster?". lecture at the Royal Albert Hall (1933)

[5] Orr, W.M.F.: The stability or instability of the steady motions of a perfect liquid and of a viscous liquid. Part 1: A perfect liquid; Part 2: A viscous liquid. Proc. R. Ir. Acad. A **27**, 9–6869138 (1907)

[6] Sommerfeld, A.: Ein Beitrag zur Hydrodynamischen Erklärung der Turbulenten Flüssigkeitbewegungen. In: Atti del IV. Congresso Internazionale dei Matematici, vol. III. Rome, pp. 116–124 (1908)

[7] Theofilis, V., Hein, S., Dallman, U.: On the origins of unsteadiness and three-dimensionality in a laminar separation bubble. Phil. Trans. R. Soc. Lond. A **358**(1777), 3229–3246 (2000)

[8] Hasselmann, K.: PIPs and POPs: The reduction of complex dynamical system-susing principal interaction and oscillation patterns. J. Geophys. Research **93**, 11015–11021 (1988)

[9] Penland, C.: Random forcing and forecasting using principal oscillation pattern analysis. Mon. Wea. Rev. **117**, 2165–2185 (1989)

[10] Mezić, I.: Spectral properties of dynamical systems, model reduction and decompositions. Nonlinear Dynamics **41**, 309–325 (2005)

[11] Schmid, P.J.: Dynamic mode decomposition of numerical and experimental data. J. Fluid Mech. **656**, 5–28 (2010)

[12] Frederiksen, J.S.: Singular vectors, finite-time normal modes, and error growth during blocking. J. Atmos. Sci **57**, 312–333 (2000)

[13] Frederiksen, J.S., Branstator, G.: Seasonal variability of teleconnection patterns. J. Atmos. Sci **62**, 1346–1365 (2005)

[14] Quinn, C., Harries, D., O'Kane, T.J.: Dynamical analysis of a reduced model for the north atlantic oscillation. J. Atmos. Sci. **78**, 1647–1671 (2021)

[15] Corazza, M., Kalnay, E., Patil, D.J., Yang, S.-C., Morss, R., Cai, M., Szunyogh, I., Hunt, B.R., Yorke, J.A.: Use of the breeding technique to estimate the structure of the analysis "errors of the day". Nonlinear Processes in Geophysics **10**, 1–11 (2003)

[16] Harries, D., O'Kane, T.J.: Applications of matrix factorization methods to climate

data. Nonlin. Processes Geophys. **27**(3), 453–471 (2020)

[17] Lumley, J.L.: The structure of inhomogeneous turbulence, pp. 166–178. Atmosphere Turbulence and Wave Propagation, Moscow: Nauka (1967)

[18] Noack, B.R., Afanasiev, K., Morzynski, M., Tadmor, G., Thiele, F.: A hierarchy of low dimensional models for the transient and post transient cyclinder wake. J. Fluid Mech. **497**, 335–363 (2003)

[19] Cordier, L., Noack, B.R., Tissot, G., Lehnasch, G., Delville, J., Balajewicz, M., Daviller, G., Niven, R.K.: Identification strategies for model-based control. Exp. Fluids **54**, 1580 (2013)

[20] Perret, L., Collin, E., Delville, J.: Polynomial identification of pod based low-order dynamical system. Journal of turblence **7**(17) (2006)

[21] Kumar, N.: Data-driven flow modelling using machine learning and data assimilation approaches. PhD thesis, Université de Poitiers (2021)

[22] Niven, R.K., Mohammad-Djafari, A., Cordier, L., Abel, M., Quade, M.: Bayesian Identification of Dynamical Systems. Proceedings. In: 39th International Workshop on Bayesian Inference and Maximum Entropy Methods In, vol. 33, pp. 1–7 (2019)

[23] Frederiksen, J.S., O'Kane, T.J.: Markovian inhomogeneous closures for rossby waves and turbulence over topography. Journal of Fluid Mechanics **858**, 45–70 (2019)

[24] Kalnay, E.: Atmospheric Modeling, Data Assimilation and Predictability. Cambridge University Press, Cambridge (2003)

[25] Gill, A.E.: Atmosphere-Ocean Dynamics. Academic Press, London (2003)

[26] O'Kane, T.J., Sandery, P.A., Kitsios, V., Sakov, P., Matear, R.J., Chamberlain, M.A., Collier, M.A., Fiedler, R., Chapman, C., Moore, T.S., Sloyan, B.: Cafe60v1: A 60-year large ensemble climate reanalysis. part i: System design, model configuration and data assimilation. J. Clim. **34**(13), 5153–516960 (2021)

[27] Roemmich, D., Coauthors: The argo program: Observing the global ocean with profiling floats. Oceanography **22**, 34–43 (2009)

[28] O'Kane, T.J., Sandery, P.A., Kitsios, V., Sakov, P., Matear, R.J., Chamberlain, M.A., Squire, D.T., Collier, M.A., Chapman, C., Fiedler, R., Harries, D., Moore, T.S., Richardson, D., Risbey, J.S., Schroeter, B.J.E., Schroeter, S., Sloyan, B., Tozer, C., Watterson, I.G., Black, A., Quinn, C.: Cafe60v1: A 60-year large ensemble climate reanalysis. part i: System design, model configuration and data assimilation. J. Clim. **34**(13), 5171–519460 (2021)

[29] Bishop, C.H., Etherton, B., Majumdar, S.J.: Adaptive sampling with the ensemble transform kalman filter. part i: theoretical aspects. Mon. Wea. Rev. **129**, 420–436 (2001)

[30] Sakov, P.: EnKF-C user gruide. CoRR **abs/1410.1233v8** (2019)

[31] Delworth, T.L., Broccoli, A.J., Rosati, A., Stouffer, R.J., Balaji, V., Beesley, J.A., Cooke, W.F., Dixon, K.W., Dunne, J., Dunne, K.A., Durachta, J.W., Findell, K.L., Ginoux, P., Gnanadesikan, A., Gordon, C.T., Griffies, S.M., Gudgel, R., Harrison, M.J., Held, I.M., Hemler, R.S., Horowitz, L.W., Klein, S.A., Knutson, T.R., Kushner, P.J., Langenhorst, A.R., Lee, H.-C., Lin, S.-J., Lu, J., Malyshev, S.L., Milly, P.C.D., Ramaswamy, V., Russell, J., Schwarzkopf, M.D., Shevliakova, E., Sirutis, J.J., Spelman, M.J., Stern, W.F., Winton, M., Wittenberg, A.T., Wyman, B., Zeng, F., Zhang, R.: GFDL's CM2 Global Coupled Climate Models. Part I: Formulation and Simulation Characteristics. J. Climate **19**(5), 643–674 (2006)

[32] Kitsios, V., Sandery, P.A., O'Kane, T.J., Fiedler, R.: Ensemble kalman filter parameter estimation of ocean optical properties for reduced biases in a coupled general circulation model. Journal of Advances in Modeling Earth Systems **13**(2), 2020–002252 (2021)

[33] Hussain, A.K.M.F., Reynolds, W.C.: The mechanisms of an organised wave in turbulent shear flow. J. Fluid Mech. **41**(2), 241–258 (1970)

[34] Bjerknes, J.: Atmospheric teleconnections from the equatorial pacific. Monthly Weather Review **97**(3), 163–172 (1969)

[35] Cordier, L., Bergmann, M.: Proper Orthogonal Decomposition: an overview. In: Lecture Series 2008 on Post-processing of Experimental and Numerical Data (2008). Von Karman Institute for Fluid Dynamics

[36] Kitsios, V., Collier, M.A., O'Kane, T.J.: Coherent Structures in the Equatorial Under Current of the Tropical Pacific Ocean. In: 23^{rd} Australasian Fluid Mechanics Conference (2022). University of Sydney

[37] Salmon, R.: Lectures on Geophysical Fluid Dynamics vol. 15. Oxford University Press, London (1998)

[38] Vallis, G.K.: Atmospheric and Oceanic Fluid Dynamics Fundamentals and Large-Scale Circulation, 2nd edn. Cambridge University Press, Cambridge (2017)

[39] Roquet, F., Madec, G., Brodeau, L., Nycander, J.: Defining a simplified yet "realisitic" equation of state for seawater. J. Phys. Oceanogr. **45**, 2564–2579 (2015)