Quantifying the Predictive Skill in Long-Range Forecasting. Part I: Coarse-Grained Predictions in a Simple Ocean Model

DIMITRIOS GIANNAKIS * AND ANDREW J. MAJDA
Courant Institute of Mathematical Sciences, New York University, New York, New York

ABSTRACT
An information-theoretic framework is developed to assess the long-range coarse-grained predictive skill in a perfect-model environment. Central to the scheme is the notion that long-range forecasting involves regimes; specifically, that the appropriate initial data for ensemble prediction is the affiliation of the system to a coarse-grained partition of phase space representing regimes. The corresponding ensemble prediction probabilities, which are computable using ergodic signals from the model, are then used to quantify through relative entropy the information beyond climatology in the partition. As an application, we study the predictability of circulation regimes in an equivalent barotropic double-gyre ocean model using a partition algorithm based on $K$-means clustering and running-average coarse-graining. Besides the established rolled up and extensional phases of the eastward jet, optimal partitions for triennial-scale forecasts feature a jet configuration dominated by the second empirical orthogonal function (EOF) of the streamfunction, as well as phases in which the jet interacts with eddies in higher EOFs. Due to mixing dynamics, the skill beyond three-state models is lost for forecast lead times longer than three years, but significant skill remains in the energy and the leading principal component of the streamfunction for septennial forecasts.

1. Introduction
Long-range forecasting is a promising emerging field in climate science, blending aspects of initial-value weather prediction and climate-change projections (Latif et al. 2006; Hurrell et al. 2009; Meehl et al. 2009; Solomon et al. 2009). Owing to significant advances in the development of general circulation models (GCMs), initial-value ensemble predictions are now possible for the coupled atmosphere-ocean system for 10-year or longer lead times (Smith et al. 2007; Keenlyside et al. 2008; Pohlmann et al. 2009). These prediction horizons are sufficiently small for the anthropogenically-forced component of climate change to be small compared to natural variability (Branstator and Teng 2010), but also long enough for that variability to have extensive socio-economic impacts.

In classical studies on decadal prediction (Boer 2000, 2004; Collins 2002), ensemble experiments are performed using one or more GCMs initialized by perturbed initial conditions relative to a reference state, and predictive skill is measured by comparing the root mean square difference (RMSD) of ensemble trajectories to its equilibrium value. However, skill metrics of this type have the drawback of not being invariant under invertible transformations of the prediction variables, and not taking into account the important issue of model error. Further challenges concern the choice of initial conditions of the ensemble members. The formulation of mathematical strategies to assess long-range predictive skill and model error in an objective manner is an open research problem (DelSole and Tippett 2007; Reichler and Kim 2008; Meehl et al. 2009; Majda and Gershgorin 2010), and forms the topic of the present work. Our analysis and results are presented in a two-paper series, in which the present paper develops strategies for phase-space coarse-graining to reveal long-range predictability in a perfect model, and Paper II (Giannakis and Majda 2011) studies quantification of model error in coarse-grained models of regime dynamics.

Throughout, we consider prediction horizons lying in the ‘confluence’ 1–10-year range between initial-value and forced boundary condition problems (Hurrell et al. 2009; Meehl et al. 2009; Solomon et al. 2009; Branstator and Teng 2010). Here the fundamental intuition is that (i) unlike weather forecasting, the forecast lead time $\tau$ is long enough for chaotic mixing to render details of the initial conditions unimportant; (ii) $\tau$ is sufficiently short so that internal variability dominates over the forced response. Item (i) means that ensemble forecasts can be made via probability distributions of large-scale observables conditioned on some coarse-grained knowledge of the initial state of the system; examples of observables are the low-frequency principal components (PCs), the amount of energy in some dynamical fields of the system, surface temperature, etc. Moreover, in light of (ii), it is appropriate to measure predictive
skill relative to the equilibrium statistics of the present-day climate. As the forecast lead time grows, the prediction probability distributions relax towards the invariant measure, or to some cyclo-stationary measure if seasonal effects are important. Nevertheless, the skill, or information content, of these distributions beyond climatology can remain large for substantially longer times than those suggested by the low-frequency Fourier components of individual trajectories (Teng and Branstator 2010).

Coarse-graining phase-space to reveal long-range predictability, not involving details of the initial data, is challenging even in a perfect-model setting. Here we demonstrate that significant progress can be made by adopting the viewpoint that (i) the dynamical systems arising in atmosphere-ocean science (AOS) are dominated on some large spatio-temporal scale by switching between different regimes in phase space (Charney and DeVore 1979; Reinhold and Peterhamburt 1982; Tung and Rosenthal 1985; Cehelsky and Tung 1987; Ghil and Robertson 2002; Dijkstra and Ghil 2005; Majda et al. 2006; Franzke et al. 2008, 2009); (ii) the appropriate coarse-grained initial data for long-range forecasting is the affiliation of the system to a partition of phase space representing the regimes. The key advantage of the above is that the phase-space partition may be constructed by data-clustering equilibrium realizations of ergodic dynamical systems without having to specify ensemble initial data. Moreover, prediction probabilities conditioned on the regimes can be evaluated empirically without having to invoke additional assumptions (e.g., Gaussianity), since detailed initial conditions are not needed to sample these distributions.

The availability of empirical cluster-conditional probabilities opens the possibility of using information theory (Leung and North 1990; Schneider and Griffies 1999; Kleeman 2002; Majda et al. 2002, 2005; DelSole 2004, 2005; DelSole and Tippett 2007; Majda and Gershgorin 2010; Teng and Branstator 2010) to characterize the additional information content beyond model climatology of the regime-conditional distributions. The natural information-theoretic functional to measure this additional information is relative entropy, which induces a notion of distance $\mathcal{D}_\tau$ between the cluster-conditional and equilibrium distributions. A related analysis, which will be undertaken in Paper II, leads to a measure of model error that is given by the distance $\mathcal{E}_\tau$ between the cluster-conditional distributions in the perfect model and a biased coarse-grained model (e.g., a Markov model). The $\mathcal{D}_\tau$ and $\mathcal{E}_\tau$ metrics both depend on the forecast lead time, $\tau$, but are invariant under invertible nonlinear transformations of observables (Majda et al. 2002).

As a concrete application of our theoretical framework, we study long-range predictability in the equivalent barotropic, double-gyre model of McCalpin and Haidvogel (1996). This simple model of ocean circulation has non-trivial low-frequency dynamics, characterized by infrequent transitions between meandering, moderate-energy, and extensional configurations of the eastward jet (analogous to the Gulf Stream in the North Atlantic). The algorithm employed here for phase-space partitioning involves building a multi-time family of clusters, computed for different temporal intervals of coarse graining; a recipe similar to kernel density estimation methods (Silverman 1986). We find that knowledge of cluster affiliation in the computed partitions carries significant information beyond climatology about the total energy and the leading two PCs (which are natural variables for the low-frequency dynamics of this system) for five- to seven-year forecast lead times, i.e., for a timescale about a factor of five longer than the maximum decorrelation time of the PCs. For lead times less than 3–5 years, our method is able to identify additional jet configurations dominated by the 2nd EOF, which are not part of the standard three-state phenomenology, but carry significant information beyond climatology. However, due to mixing dynamics, the additional information is lost when making three-year or longer forecasts.

The plan of this paper is as follows. In §2, following a brief review of relevant concepts of information theory, we lay out the general principles of our framework for assessing skill in long-range forecasting. These principles are specialized in §3 with reference to the running-average phase-space partitioning scheme. Section 4 is our application section, where we introduce the 1.5-layer ocean model, use the phase-space partitioning scheme of §3 to identify its circulation regimes, and apply the information-theoretic framework of §2 to study the intrinsic predictive skill associated with these regimes. We conclude in §5.

2. Information theory and long-range forecasting

a. Setting

We consider a high-dimensional, strongly chaotic, ergodic, and mixing dynamical system,

$$\frac{dx}{dt} = F(x), \quad x(t) \in \mathbb{R}^m, \quad m \gg 1,$$

observed through a reduced set of variables,

$$z(t) = G(x(t)),$$

with $z(t) \in \mathbb{R}^n$ and $n \ll m$. For instance, in §4 $x(t)$ will be the streamfunction $h(\mathbf{r}, t)$ of the 1.5-layer model (29) and $z(t)$ the leading 20 corresponding PCs. Ergodicity means that if $A(x(t))$ is a function of the state vector (e.g., the energy $E$), then empirical temporal averaging can be used to approximate expectation values associated with the equilibrium measure $p_{eq}(x)$, i.e.,

$$\frac{1}{\delta t} \sum_{i=0}^{s-1} A(x(t - i \delta t)) \approx \int dx p_{eq}(x) A(x)$$

(3)
for a large-enough number of samples s and (for simplicity)
uniform sampling interval δt. Throughout this work we re-
strict attention to stationary dynamical systems, but the
methods for quantifying predictive skill presented below
(and the corresponding formalism for model error de-
veloped in Paper II) can be generalized to deal with time-
periodic equilibrium statistics (Majda and Wang 2010).

Broadly speaking, the theory for long-range forecasting
developed here associates dynamical regimes with a parti-
tion of observation space \( \mathbb{R}^n \) into \( K \) clusters,
\[
\Xi = \{\xi_1, \ldots, \xi_K\}, \quad \xi_k \subset \mathbb{R}^n,
\] (4)
such that whenever \( z(t) \) belongs in \( \xi_k \), the corresponding
values of \( A(t) \) exhibit a coarse-grained degree of similarity
by some measure. The information content in \( \Xi \) for long-
range forecasting (below, sometimes referred to loosely as
the ‘skill’ of \( \Xi \) range forecasting (below, sometimes referred to loosely as
such that whenever \( p = p' \), positive when \( p \neq p' \), and invariant under invertible
changes of variable, it describes a ‘distance’ between
probability distributions (though it is not symmetric in its
arguments, and it does not obey the triangle inequality).

In particular, \( \mathcal{P}(p, p') \) may be interpreted alternatively as
(i) the gain of information about \( A \) achieved by a mea-
surement whose posterior distribution is \( p \), when the dis-
tribution of \( A \) prior to the measurement was \( p' \); (ii) the
lack information of \( p' \) about \( A \), when realizations of \( A \) are actually generated by \( p \).

Let \( p_r(A) = P(A(t + \tau) \mid S(t)) \) be a probability dis-
tribution for observable \( A \) at forecast lead time \( \tau \geq 0 
conditioned on some variable \( S \) measured at time \( t \), and
let \( p_{\text{eq}}(A) = \lim_{\tau \to \infty} p_r(A) \) be the corresponding equilib-
rium measure. In general, \( S \) may be any variable whose
knowledge provides provides future information about \( A 
but here we explicitly consider that \( S(t) \in \{1, \ldots, K\} \) is
the cluster-affiliation sequence associated with a partition
\( \Xi \) in (4); i.e., \( S(t) = k \) means that the observation vector
\( z(t) \) belongs in cluster \( \xi_k \in \Xi \). As described in §3, \( S(t) \) is
uniquely determined given initial data
\[
Z(t) = \{z(t), z(t - \delta t), \ldots, z(t - (q - 1) \delta t)\},
\] (6)
spanning a running-average window \( \Delta \tau = (q - 1) \delta t \). We
use
\[
p^k_r(A) = p(A(t + \tau) \mid S(t) = k), \quad k \in \{1, \ldots, K\},
\] (7)
to denote this family of time-dependent, cluster-conditional probabilities for \( A \).

By interpretation (i) of relative entropy, the quantity
\[
\mathcal{D}^k_r = \mathcal{P}(p^k_r, p_{\text{eq}})
\] (8)
measures the additional information beyond climatology
associated with regime \( \xi_k \in \Xi \). That is, the metric (8)
quantifies the skill of ensemble predictions when it is known
that the system visited cluster \( k \) at time \( \tau = 0 \). Denoting
by
\[
\pi_k = p(S(t) = k)
\] (9)
the equilibrium probability of affiliation with cluster \( k \), the
weighted mean
\[
\mathcal{D}_\tau = \sum_{k=1}^{K} \pi_k \mathcal{D}^k_r
\] (10)
is a superensemble-type measure of predictive skill. That is, $\mathcal{D}_\tau$ gives the expected gain of information beyond climatology associated with $\Xi$ for initial data drawn from the equilibrium measure.

As discussed by DelSole (2004, 2005), $\mathcal{D}_\tau$ has an additional interpretation as the mutual information between $S(t)$ and $A(\tau + t)$, viz.

$$I(A(\tau + t); S(t)) = \sum_{S(t) = 1}^K \int dA(\tau + t) \, p(A(\tau + t), S(t)) \times \log \frac{p(A(\tau + t), S(t))}{p(A(\tau + t))p(S(t))} = \mathcal{D}_\tau(A),$$

where the last equality follows by Bayes’ theorem, $p(A(\tau + t), S(t)) = p_k(A)^T p_k$. The above is also equal to the transinformation measure introduced by Leung and North (1990), as well as the expectation value with respect to $p_k$ from (9) of the predictive information of Schneider and Griffies (1999). In particular, the classical result that $I(A(\tau + t); S(t))$ vanishes if and only if $A(\tau + t)$ and $S(t)$ meet the statistical-independence condition (Cover and Thomas 2006, Chap. 2)

$$p(A(\tau + t), S(t)) = p(A(\tau + t))p(S(t)) = p_{eq}(A)p_{eq}(S),$$

intuitively links with the notion that the predictive information content of the partition should vanish in the infinitely far future, when $A(\tau + t)$ relaxes to equilibrium. Because relative entropy is unbounded from above, it is convenient to convert $\mathcal{D}_\tau$ to a predictability score $\delta_\tau \in [0, 1]$ via the transformation

$$\delta_\tau = 1 - \exp(-2\mathcal{D}_\tau).$$

In certain cases involving normally distributed variables, the value of $\delta_\tau$ is equivalent to a standard correlation measure (Joe 1989; DelSole 2005).

The skill metric in (10) can be used to choose among a collection of partitions $\{\Xi_1, \Xi_2, \ldots\}$ the optimal partition for observable $A$, even if the candidate partitions have different cluster numbers, were computed via different clustering algorithms, or different observation spaces were used. What $\mathcal{D}_\tau$ does not reflect, however, is the lack of predictive information in $\Xi$ compared with a more fine-grained partition, $\Xi'$, consisting of $L > K$ clusters, such that each of the elements of $\xi'_k \in \Xi'$ is a subset of one and only one $\xi_k \in \Xi$. Here, given the same history $Z(t)$ of initial data, the coarse-grained distributions $p_k^\xi(A)$ ‘project down’ some of the extra information contained in $p_k^\xi(A)$, and as a result some loss of skill is expected to occur.

By interpretation (ii) of the relative entropy, the information loss incurred by using $p_k^\xi$ instead of $p_k^\tau$ is measured by $\mathcal{E}_\tau^\xi = \mathcal{P}(p_k^\tau, p_k^\xi)$. As above, $\mathcal{E}_\tau^\xi$ may be aggregated in the superensemble-type measure $\mathcal{E}_\tau^\xi = \sum_{l=1}^L \pi_l^\xi \mathcal{E}_\tau^\xi_l$, with $\pi_l^\xi = p(S_l' = l)$. The latter measures the mean loss of predictive information about $A$ due to the coarse-graining operation $\Xi' \mapsto \Xi$. For short prediction times, where details in the initial data play a more important role, $\mathcal{E}_\tau^\xi$ may be significant. If, however, $\tau$ is long enough so that fine-grained memory of $Z(t)$ is lost due to mixing, then $\Xi$ is expected to be a good surrogate for $\Xi'$, in the sense of $\mathcal{E}_\tau^\xi$ being small.

As expected intuitively, the coarse graining of $\Xi'$ can only decrease or preserve its predictive information content about an observable. This is a direct consequence of the data-processing inequality (DelSole 2005; Cover and Thomas 2006) in information theory, which states that $I(Y; X) \geq I(Z; X)$ for random variables that meet the Markov condition

$$p(X, Y, Z) = p(Z \mid Y)p(Y \mid X)p(X).$$

The above is satisfied for $\{A(\tau + t), S'(t), S(t)\}$ in that order, since $p(S(t) \mid S'(t), A(\tau + t)) = p(S(t) \mid S'(t))$ holds trivially when coarse graining $\Xi'$ to form $\Xi$. Similar lower bounds for predictability can be derived for all observables $B(\tau + t)$ which are sufficient statistics of $A(\tau + t)$, i.e., $p(A(\tau + t) \mid B(\tau + t), S(t)) = p(A(\tau + t) \mid B(\tau + t))$. Here, the Markov chain is formed by $\{S(t), B(\tau + t), A(\tau + t)\}$, leading to the bound $\mathcal{D}_\tau(B) \geq \mathcal{D}_\tau(A)$.

Before closing this section, we note that a key assumption of our analysis has been that the PDFs in (7) and (9) can be evaluated without error, other than error due to finite sample size. Of course, this is a very strong assumption, which is unlikely to be satisfied in realistic AOS applications. In Paper II we will relax that assumption, and study the loss of information occurring when a model $M$ produces approximations $p_k^M$ of the cluster-conditional prediction probabilities that are systematically biased away from $p_k^\tau$. For that purpose we shall make use of a relative-entropy error measure $\mathcal{E}_k^\tau = \mathcal{P}(p_k^\tau, p_k^M)$ with a similar mathematical structure to $\mathcal{E}_k^\tau$ introduced above. Note, however, a crucial difference between $\mathcal{E}_\tau^\xi$ and $\mathcal{E}_\tau$: The former measures the loss of information due to coarse-graining in an otherwise unbiased model. In contrast, the error measures developed in Paper II reflect a situation where using the model prediction PDFs directly in $\mathcal{D}_\tau$ can convey false predictive skill.

3. Running-average coarse-grained prediction

So far, our discussion has focused on general theoretical aspects of phase-space coarse-graining, which to a large extent are independent of the algorithm used to partition the observed data. In this section, we develop a practical algorithm for partitioning data sets generated by ergodic dynamical systems. The first step of the approach is a training stage ($\S 3a$), where the input data set is partitioned into clusters by means of a kernel-smoothed variant of $K$-means clustering. This is followed by a prediction
stage (§3b), where the cluster coordinates computed in the training stage are used to produce the partition $\Xi$ in (4) and the corresponding prediction probabilities $p_k^\tau$ in (7) for observables of interest. Both training and prediction stages involve running averages, respectively for coarse-graining the training time series and the data for prediction; a recipe similar to kernel density-estimation methods (Silverman 1986). Throughout, we restrict attention to stationary systems, but our information-theoretic approach can be generalized to nonstationary problems with time-periodic equilibrium statistics by casting the periodic system into a time-independent skew-symmetric-product form (Majda and Wang 2010; Gershgorin and Majda 2010).

a. Training stage

The training stage of our coarse-grained modeling procedure involves taking a data set

$$\mathcal{Z} = \{z(s-1)\delta t), z(s-2)\delta t), \ldots, z(0)\}, \quad (15)$$

consisting of $n$-variate observations $z(t)$ from (2) of the high-dimensional ergodic dynamical system in (1) over a time interval $[0, (s-1)\delta t] = [0, T]$, and partitioning it into $K$ pairwise-disjoint clusters, which are characterized by a set of coordinates

$$\Theta = \{\theta_1, \ldots, \theta_K\}, \quad \theta_k \in \mathbb{R}^n. \quad (16)$$

Associated with $\Theta$ is an integer-valued cluster-affiliation sequence $\Gamma(t) \in \{1, \ldots, K\}$, such that $\Gamma(t) = k$ means that the observation vector $z$ at time $t \in [0, T]$ belongs in the $k$-th cluster. Note that $\Gamma(t)$ is not the same affiliation function as the $S(t)$ used in §2b; $\Gamma(t)$ provides a historical identification of the regimes in the input time series $z(t)$, whereas $S(t)$ is computed on-the-fly using data observed at $t \notin [0, T]$. In the ocean application of §4, $z(t)$ will be the leading $n = 20$ PCs of the streamfunction,

$$z(t) = (PC_1(t), \ldots, PC_n(t)) \in \mathbb{R}^n, \quad (17)$$

sampled with a uniform time step $\delta t = 20$ days.

A major challenge in evaluating affiliations to clusters is incorporating a notion of temporal regularity in $\Gamma(t)$ (Horenko 2009, 2010). This is a natural requirement for data generated by physical dynamical systems, but as has been pointed out (Christiansen 2007), is frequently not taken into consideration in cluster analysis of AOS data sets. Here, we regularize $\Gamma(t)$ by smoothing our input data by running averaging—a widely used method for preprocessing AOS data sets (Ghil and Robertson 2002). Specifically, we fix a temporal window $\Delta t = (q'-1)\delta t$, and replace $z(t)$ by

$$z^{\Delta t}(t) = \sum_{i=1}^{q'} z(t - (i-1)\delta t)/q'. \quad (18)$$

The cluster coordinates $\theta_k$ are then determined by $K$-means clustering (MacQueen 1967) using the coarse-grained time series $z^{\Delta t}(t)$ as input data. That is, the $\theta_k$ are computed via standard iterative algorithms (Duda et al. 2000) by minimizing with respect to $\Theta$ from (16) the error functional

$$L(\Theta) = \sum_{k=1}^{K} \sum_{i=q'-1}^{s-1} \gamma_k(i\delta t)\|z^{\Delta t}(i\delta t) - \theta_k\|^2, \quad (19)$$

where

$$\gamma_k(t) = \begin{cases} 1, & k = \text{argmin}_j \|z^{\Delta t}(t) - \theta_j\|_2, \\ 0, & \text{otherwise}, \end{cases} \quad (20)$$

is the weight of the $k$-th cluster at time $t = i\delta t$, and $\|v\|_2 = (\sum_{i=1}^{n} v_i^2)^{1/2}$ denotes the Euclidean norm. The resulting cluster-affiliation sequence is then given by

$$\Gamma(t) = \underset{k}{\text{argmax}} \gamma_k(t). \quad (21)$$

The minimization of $L$ is to be repeated for different values of $\Delta t$, giving a one-parameter family of cluster coordinates $\Theta^{\Delta t}$ at various coarse-graining levels of the training data set. Effectively, the running-average interval $\Delta t$ plays the role of a smoothing parameter in kernel methods (Silverman 1986). That is, the convolution operation in (18) suppresses the high-frequency components of $z(t)$, leading to a reduction of the number of transitions in $\Gamma(t)$ with $\Delta t$ (see Fig. 5).

b. Prediction stage

We now describe how the cluster coordinates $\Theta^{\Delta t}$ of §3a induce a $K$-element partition $\Xi$ from (4) of observation space, and how the corresponding prediction probabilities $p_k^\tau(A)$ in (7) can be estimated by conditional binning time series of observables $A(t)$. A key element of this procedure is to establish a rule for assigning cluster affiliation, $S(t)$, given data $z(t)$ that are not part of the training data set $\mathcal{Z}$ from (15) (i.e., here $t \notin [0, T]$). For this purpose, we introduce an additional coarse-graining interval, $\Delta \tau = (q-1)\delta t$, and use it to process $z(t)$ via the formula

$$z^{\Delta \tau}(t) = \sum_{i=1}^{q} z(t - (i-1)\delta t)/q. \quad (22)$$

The online cluster affiliation at time $t$ is then determined by

$$S(t) = \underset{k}{\text{argmin}} \|z^{\Delta \tau}(t) - \theta_k^{\Delta t}\|_2. \quad (23)$$

Therefore, $S(t)$ depends on both $\Delta t$ and $\Delta \tau$, and, as stated in §2b, is uniquely determined by the trajectory $Z(t)$ in (6) over the time interval $[t - \Delta \tau, t]$. Associated with the affiliation rule in (23) are the subsets of observation space

$$\xi_k = \{z(t) \in \mathbb{R}^n : S(t) = k\}, \quad (24)$$
which together make up a natural pairwise-disjoint partition \( \mathcal{Z} \) of the form given in (4).

The remaining ingredients needed to measure the information content in \( \mathcal{Z} \) for long-range forecasting are the prediction probabilities \( p^k(A) \). By virtue of ergodicity, these probabilities may be estimated as follows. First, obtain a sequence of observations \( z(t) \) (independent of the training data set \( Z \)) and the corresponding time series \( A(t) \) of the prediction observable. Second, using (23), compute the affiliation sequence \( S(t) \). For given prediction horizon \( \tau \), and for each \( k \in \{1, \ldots, K\} \), collect the values

\[
A^k = \{A(t + \tau) : S(t) = k\},
\]

Then, set distribution bin boundaries \( A_0 < A_1 < \ldots \), and compute the occurrence frequencies

\[
\hat{p}^k(A_i) = N_i / N,
\]

where \( N_i \) is the number of elements of \( A^k \) lying in \([A_{i-1}, A_i]\), and \( N = \sum_i N_i \). Note that the \( A_i \) are vector-valued if \( A \) is multi-variate. By ergodicity, in the limit of an infinite number of bins and samples, the estimators \( \hat{p}^k(A_i) \) converge to the continuous probability density functions (PDFs) \( p^k(A) \) in (7). The equilibrium PDF \( p_{eq}(A) \) and the cluster-affiliation probabilities \( \pi_k \) from (9) may be evaluated in a similar manner. Together, the estimates for \( p^k \), \( p_{eq} \), and \( \pi_k \) are sufficient to implement the skill metrics of §2. In particular, if \( A \) is a scalar observable (as will be the case in §4), the relative entropies in (8) can be computed by standard one-dimensional quadrature.

We now come to an important point about the prediction probabilities from (7): Because \( p^k(A; \Delta T) \) are evaluated independently for each pair \( \Delta T = (\Delta t, \Delta \tau) \) of coarse-graining intervals, there is no reason why one should use the same \( p^k(A; \Delta T) \) for all forecast lead times. In particular, given a collection \( \{\Delta T_1, \Delta T_2, \ldots\} \) of coarse-graining parameters, the natural prediction probabilities to use are the ones that maximize the expected predictive skill (10), viz.

\[
p^{*k}_r(A) = p^k_r(A; \Delta T_i), \quad i = \arg\max_j D_r(A; \Delta T_j),
\]

with corresponding optimal super-ensemble predictive skill

\[
D^*_r(A) = D_r(A; \Delta T_i), \quad \delta^*_r(A) = 1 - \exp(-2D^*_r(A)).
\]

We will see in §4 that the \( p^{*k}_r \) have significantly higher skill than the individual prediction probabilities \( p^k_r \) for the energy and leading PCs in the 1.5-layer model.

4. Demonstration in a double-gyre ocean model

We now apply the theory and methods developed in §§2,3 to study the long-range predictability of circulation regimes in the reduced-barotropic double-gyre model of McCalpin and Haidvogel (1996). Following a brief introduction to the 1.5-layer model (§4a), we demonstrate in §4b that partitions of observation space constructed via running-average \( K \)-means clustering, and used in conjunction with the procedure of §3b for computing empirical prediction probabilities, retain information \( D_t \) beyond climatology about large-scale observables (e.g., the energy and the leading streamfunction PCs) for lead times up to \( \tau \simeq 7 \) years, or \( \simeq 6.5 \) multiples of the longest decorrelation time of the PCs (Fig. 4).

Via the information-theoretic skill measures of §2, we study systematically the interplay between coarse-graining the training time series and/or the interval for initial cluster affiliation, and how that affects predictive skill. We find that, for suitable choices of running-average windows, seven-state partitions contain significant information beyond the standard three-state description of regime behavior in this class of ocean models (Berkhoff and McWilliams 1999; McCalpin and Haidvogel 1996), but, due to mixing dynamics, that additional information is lost for predictions beyond \( \tau \simeq 1000 \) days. Despite the eventual loss of predictive information content, the circulation regimes in \( K = 7 \) models (Figs. 8 and 9) have rich spatial structure, including a configuration of the eastward jet dominated by the second EOF, as well as interactions of the jet with mid-basin eddies.

\[a. The ocean model\]

The so-called 1.5-layer model (McCalpin and Haidvogel 1996) describes the dynamics of wind-driven ocean circulation as the motion of two immiscible, vertically-averaged layers of fluid of different density under the influence of wind-induced shear, Coriolis force (in the \( \beta \)-plane approximation), and subgrid-scale diffusion. The lower layer is assumed to be infinitely deep and at rest, whereas the upper layer is governed by a quasigeostrophic equation for the streamfunction \( h \) (which, in this case is equal to the interface displacement), viz.

\[
(\Delta - \gamma^2) \partial_t h = -\frac{g^*}{f_0} \nabla^\perp h \cdot \nabla \Delta h - \beta \partial_x h
\]

\[
- r \Delta h - A_b \Delta^3 h + \frac{f_0}{\rho g^* H^*} \text{curl} \tau.
\]

In the above, \( (x, y) = \mathbf{r} \) respectively are the zonal and meridional directions, \( \nabla = (\partial_x, \partial_y) \) is the gradient operator, \( \nabla^\perp = (-\partial_y, \partial_x) \) is the orthogonal gradient (giving the velocity vector via \( \mathbf{v} = g f_0^{-1} \nabla^\perp h \)), and \( \Delta = \partial_x^2 + \partial_y^2 \) is the Laplace operator. The model parameters are defined in Table 1. The kinetic and potential energies, respectively given by \( E_{\text{kin}} = \rho H^* \int dr \| \mathbf{v}(r, t) \|^2 / 2 \) and \( E_{\text{pot}} = \rho g^* \int dr h^2(r, t) / 2 \), make up the total energy, \( E = E_{\text{kin}} + E_{\text{pot}} \), which will be one of our main prediction observables.
We adopt throughout the parameter values in §2 of McCalpin and Haidvogel (1996), as well as their canonical asymmetric double-gyre wind forcing. With this forcing, the 1.5-layer model develops an eastward-flowing separating jet configuration analogous to the Gulf Stream in the North Atlantic. Moreover, the model features the essential dynamical mechanisms of equivalent barotropic Rossby waves, lateral shear instability, and damping.

The model was integrated by Rafail Abramov using a pseudospectral code on a 180 × 140 uniform grid of size \( \Delta r = 20 \) km, and 4th-order Runge-Kutta timestepping of waves, lateral shear instability, and damping.

In what follows, we view the solution of the 1.5-layer model as the true signal (1) from nature, i.e., we set \( x(t) = h(r,t) \). Moreover, we consider that \( x(t) \) is observed through the leading 20 PCs of the streamfunction, \( PC_i(t) = \int \text{d} r \text{EOF}_i(r) h'(r,t) \), where \( \text{EOF}_i(r) \) is the \( i \)-th empirical orthogonal function in the streamfunction metric (see the caption to Fig. 3 for a definition), and \( h'(r,t) = h(r,t) - H(r) \) is the streamfunction anomaly. Thus, the observation vector \( z(t) = G(x(t)) \) is 20-dimensional, and has the form given in (17).

A special property of the 1.5-layer model, shown in Fig. 3 is that the EOFs with the largest eigenvalues (i.e., those that carry most of the temporal variance in the streamfunction) also have the longest decorrelation time, \( \tau \). This ensures that the observation vector \( z(t) \) consisting of the leading PCs is an appropriate representation of the low-frequency dynamics, as has been noted by Berloff and McWilliams (1999). Moreover, the fact that the autocorrelation functions of \( PC_1 \) and \( PC_2 \) (Fig. 2) are positive for all lags means that \( \tau \) are appropriate time scales for measuring the decay of predictive skill. If the \( \tau \)’s did not match well with the explained variances of the corresponding EOFs (as one might expect in more comprehensive models of the ocean, involving, e.g., baroclinic modes), the problem of constructing an observation space that adequately captures regime behavior would be significantly more challenging (Franzke et al. 2008, 2009). One way of alleviating such difficulties may be to project the data to an empirical basis of predictable patterns (Schneider and Griffies 1999; Majda et al. 2002; DelSole and Chang 2003), which, rather than maximizing explained variance, optimize forecast error in a suitable metric.

The EOFs that will play major role in the coarse-grained models for regime behavior developed below are \( \text{EOF}_1 - \text{EOF}_4 \), shown in Fig. 1. In particular, \( \text{EOF}_1 \) and, to a lesser extent, \( \text{EOF}_2 \), both feature strong zonally-elongated gyres, and in each case one of the gyres is attached to the western domain boundary. The boundary regions between these counter-rotating gyres give rise to a well-defined jet separating from the western meridional boundary. Clearly, \( \text{EOF}_1 \) is related to an extensional phase, during which the jet follows a nearly straight path as it travels east. The jet in \( \text{EOF}_2 \) also flows in a predominantly easterly direction, following a sinusoid-like path with a characteristic \( O(100) \) km dip to the south at \( x \sim 750 \) km. On the other hand, the most prominent structures in EOFs 3 and 4 are more circular eddies that are detached from the western boundary. When the corresponding PCs have magnitudes comparable to \( PC_1 \) and/or \( PC_2 \), these eddies interact with the coherent jets in the leading two EOFs, affecting, for instance, their penetration length into the basin. The fourth EOF is expected to influence strongly the jet penetration
length, as it contains a prominent eddy located in the same region (at $x \sim 1500$ km latitude) as the eastern portions of the EOF$_1$ and and EOF$_2$ jets. As shown in the lower-right panel of Fig. 9, the dominant projection of the mean state $H(r)$ is to EOF$_2$ (as manifested by the similarity between the mean-state and EOF$_2$ jets), but its projection coefficients onto EOFs 1 and 4 are also appreciable.

b. Intrinsic long-range forecasting skill for coarse-grained observables

For our clustering and prediction-probability calculations we took a time series $z(t)$ from (17) of the leading $n = 20$ PCs (see Figs. 1 and 3), consisting of a total of $s = 1.6 \times 10^5$ samples taken uniformly every $\delta t = 20$ days. That is, the total observation time span is $s \delta t = 3.2 \times 10^6$ days $\approx 8767$ years. Our training data set $\mathcal{Z}$ (15) is the first half of that time series, i.e., $t \in [0, T]$, with $T = 1.6 \times 10^5$ days.

The prediction observables considered in this study are the energy $E$ and the leading-four streamfunction PCs. In light of the conventional low-, middle-, and high-energy phenomenology of the 1.5 model (McCalpin and Haidvogel 1996; Berloff and McWilliams 1999), energy is a natural observable to consider for long-range forecasting. Moreover, as we will demonstrate below, the time-averaged spatial features of the regimes are well captured by the leading four PCs. We used the portion of the time series with $t > T$ to compute the cluster-conditional time-dependent probabilities $p_{\mathbf{X}}(t)$ (7) for these observables via the procedure described in §3. Thus, the data used to estimate $p_{\mathbf{X}}(t)$ are independent of the input data to the algorithm for evaluating the cluster coordinates $\theta_k$ (16).

All prediction PDFs were estimated by binning the $s/2$ prediction samples in $n_B = 100$ bins of uniform width. The entropy integrals in the skill metric $D_\tau$ (10) were evaluated via the standard trapezoidal rule. We verified the robustness of our results against sampling and quadrature errors respectively by halving the length of the prediction time series, or the number $n_B$ of distribution bins. Neither change affected significantly the skill measures presented in Fig. 4 and Table 2. Moreover, we tested for robustness of the computed cluster-coordinates $\theta_k$ in (16) by using half of our training data. This did not impart significant changes in the spatial structure of the regimes in Figs. 7 and 8, nor the EOF projection coefficients in Fig. 9.

Following the strategy laid out in §§3a,b, we vary the running-average time intervals $\Delta t$ and $\Delta \tau$, used respectively to coarse-grain $\mathcal{Z}$ and the time series (22) of initial data, seeking to maximize (for the given choice of observable and forecast lead time $\tau$) the information content $D_\tau$ from (10) beyond climatology [or, equivalently, the skill score $\delta^*_{\tau}$ in (13)] in the resulting partition from (24) of observation space. In Fig. 4 we display a sample of the $\delta^*$ results for fixed $\Delta t = 1000$ days (i.e., a value comparable to the decorrelation time, $\tau_1 = 1165$ days, of PC$_1$), and representative values of short and long initial-data windows, respectively $\Delta \tau = 0$ and $\Delta \tau = \Delta t = 1000$ days. For the time being, we consider models with either $K = 3$ or $K = 7$ clusters, and subsequently (in §4c) study in more detail the relevance of these choices from a physical and information-theoretic standpoint.

There are a number of important points to be made about Fig. 4. First, for the chosen observables, the skill score $\delta^*_{\tau}$ (28) of the optimal partitions is significant for prediction horizons that exceed the longest decorrelation time in the $z(t)$ components used for clustering by a large margin. The fact that decorrelation times are poor indicators of intrinsic long-range predictability has been noted in other AOS applications (Teng and Branstator 2010). Here, the decay in the $\delta^*_{\tau}$ score for energy over one $e$-folding time corresponding to $\tau_1$ is $\delta^*_{\tau} / \delta^*_0 \approx 0.7$, or a factor of five weaker decay than $e^{-2} \approx 0.14$ expected for a purely exponential decay (the comparison is with $e^{-2}$ rather than $e^{-1}$ because $\delta^*_{\tau}$ is associated with squared correlations). Prediction skill for energy remains significant up to $\tau \approx 3000$ days ($\delta^*_{5000} / \delta^*_0 \approx 0.07$), or three times the decorrelation time of PC$_1$. This means that predictions approaching the decadal scale are possible for $E$, given knowledge at time $\tau = 0$ of the system’s affiliation to the regimes associated with partition $\mathcal{Z}$ in (4). Note that no fine-grained information about the initial conditions is needed to make these predictions. Uncertainty in initial conditions is a well-known obstacle in long-range forecasts (Keenlyside et al. 2008; Hurrell et al. 2009; Meehl et al. 2009; Solomon et al. 2009).

Second, as illustrated by the discrepancy between the $\delta^*$ scores evaluated for $\Delta \tau = 0$ and 1000 days, the time window $\Delta \tau$ that maximizes the information beyond climatology in the partition depends on both the observable and the forecast lead time. More specifically, in the calculations used to produce the $\delta^*_{\tau}$ versus $\tau$ curves in Fig. 4, the optimal $\Delta \tau$ for mid-term prediction ($\tau \lesssim 500$ days) of the energy is around 500 days, but that value rapidly decreases to essentially no coarse-graining ($\Delta \tau = 0$) when $\tau$ extends beyond the two-year horizon. On the other hand, $\Delta \tau = 0$ is optimal for all values of the prediction lead time $\tau$ in the case of the PCs. The larger optimal values of $\Delta \tau$ for mid-range forecasts of the energy are consistent with the fact that energy is a globally-integrated quantity, whereas the PCs describe local spatial structures that carry significant information when observed over short timescales. The fact that the optimal $\Delta \tau$ for long-range forecasting is small is beneficial from a practical standpoint, since it alleviates the need of collecting initial data over long periods.

Third, as alluded to in the beginning of this section, the $K = 7$ partitions have significantly higher skill than the $K = 3$ ones for mid-range forecasts (up to three years), but that additional skill is lost in the large lead-time regime. In particular, the $\delta^*$ skill curves of the $K = 3$ and $K = 7$ models meet at approximately $\tau = 2000$ days for $E$, 500 days
for PC₁, and 1000 days for PC₂.

A final point about Fig. 4 pertains to the non-monotonicity of \( \delta_\tau \) (equivalently, \( D_\tau \)) for \( E \). It is a general result, sometimes referred to as the generalized second law of thermodynamics, that if the dynamics of an observable are Markovian, then the corresponding relative entropy \( D_\tau \) decreases monotonically with \( \tau \) (Kleeman 2002; Cover and Thomas 2006). Thus, the increasing portion of the \( \delta_\tau(E) \) versus \( \tau \) curve for \( \Delta \tau = 700 \) days, and \( \tau \lesssim 500 \) days is a direct evidence of non-Markovianity of the energy observable. As discussed in Paper II, this has important implications for model error when the corresponding cluster-affiliation sequence is approximated by a Markov process.

c. Characterizing the optimal number of clusters

Besides quantifying the information content for prediction, the family of relative-entropy metrics \( D_\tau \) in (10) have high utility in estimating the optimal number of clusters in a coarse-grained partition \( \Xi \) from (4) of observation space. As we saw in the preceding section, \( D_\tau \) depends explicitly on the prediction horizon, \( \tau \), as well as the choice of observable \( A \). It also depends implicitly on the choice of the running-average parameters, \( \Delta t \) and \( \Delta \tau \), in the training and prediction stages of coarse-grained model building. Thus, our approach of characterizing the optimal \( K \) will generally not lead to a unique answer, but to an assessment that depends on the observable of interest, the forecast lead time, and the resolution in time at which data analysis is performed. Our viewpoint is that assigning a unique optimal \( K \), valid globally for all of the latter parameters, is too restrictive a constraint. Indeed, this type of approach has sometimes led to controversy in AOS applications (Christiansen 2007).

Here, we consider the \( \delta_\tau \) predictive skill scores for energy and the leading four PCs at lead time \( \tau = 0 \). These scores measure the skill of the partitions in (24) for marginal (i.e., univariate) classification of \( E \) and PC₁–PC₄. Restricting attention to this family of coarse-grained observables and present-day classification skill will be sufficient to expose the principles of our approach, which can then be applied for nonzero \( \tau \) and more general (possibly multivariate) observables. This analysis will also provide justification for working with \( K = 7 \) models in §4b.

In Fig. 5 we display \( \delta_\Omega \) as a function of the initial data window \( \Delta \tau \) for values of the coarse-graining interval in the training stage \( \Delta t \in \{200, 1000\} \) days and \( K \in \{3, 6, 7, 8\} \). As expected, the results in that figure exhibit the general trend that the information content in \( \Xi \) is an increasing function of \( K \). However, for certain values of \( K \) the additional information gained by further increasing the cluster number is marginal, sometimes even negligible. For each observable, we estimate the optimal number of clusters as the value of \( K \) for which that behavior occurs. For instance, the skill score for \( E \) evaluated for \( \Delta t = 1000 \) days increases significantly between \( K = 6 \) and \( K = 7 \), whereas the corresponding increase in going from \( K = 7 \) to \( K = 8 \) is markedly smaller. The same behavior is observed for PC₃ and \( \Delta \tau = 0 \). Note that PC₃ is associated with a strong mid-basin eddy (see Fig. 1), which is expected to play an important role in breaking up the jet in the low-energy regimes. Put together, these observations provide a strong indication that it is meaningful to extend the standard three-state phenomenology of McCalpin and Haidvogel (1996) to seven states, but working with \( K > 7 \) models introduces a level of detail that is not of great importance for low-frequency dynamics.

Figure 5 also illustrates the influence of the coarse-graining interval \( \Delta t \) for the training data set \( \mathcal{Z} \) (15) on the skill of the partitions; a topic that we did not touch upon in §4b. Here, the main observation is that compared with their \( \Delta t = 1000 \)-day counterparts, the partitions evaluated at coarse-graining level \( \Delta t = 200 \) days feature a substantial gain of information for PC₃, a small gain of information for PC₂, but an appreciable decrease of information in the larger-scale observables, namely \( E \) and PC₁. This is consistent with the expectation that the optimal coarse graining of the training data should be less extensive for observables that vary on short timescales. Note, however, that the optimal values of \( \Delta t \) and \( \Delta \tau \) are, in general, not equal. In Fig. 5 this is exemplified most strongly by PC₁, where the highest information content occurs for partitions evaluated with a \( \Delta t = 1000 \)-day coarse graining of \( \mathcal{Z} \), but only if the initial-data interval is short (\( \Delta \tau \lesssim 100 \) days).

d. The physical properties of the regimes

We now study the spatial and temporal properties of the regimes associated with the coarse-grained partitions of §4c. For concreteness, we focus on the models displayed in Table 2, viz. a \( K = 3 \) model with running-average windows \( (\Delta t, \Delta \tau) = (1000, 1000) \) days and two \( K = 7 \) models, respectively with \( (\Delta t, \Delta \tau) = (1000, 0) \) days and \( (\Delta t, \Delta \tau) = (1000, 1000) \) days. The parameters of the \( K = 3 \) model were motivated by the analyses of McCalpin and Haidvogel (1996) and Berloff and McWilliams (1999), which associate the meandering, mean-flow resembling, and extensional circulation regimes of the 1.5-layer model with bands of low, moderate, and high values of the energy observable. More specifically, the chosen \( \Delta \tau \) value is a reasonable compromise for simultaneously maximizing the \( D_\Omega \) skill metrics in Fig. 5 for energy and the leading two PCs.

In the case of the \( K = 7 \) partitions, the penalty in information content about the leading PCs when working with the \( \Delta \tau = O(1000) \)-day windows that are optimal for energy is significantly higher than for the three-state partitions, as manifested, e.g., by the steepness of the \( \delta_\Omega \) curves for PC₂. Therefore, the time-averaged large-scale spatial structures of the flow in the optimal \( K = 7 \) partitions from the point of view of the leading PCs do not necessarily correspond to
well-defined regimes in the energy variable, as is the case with the simpler three-state models. We illustrate this discrepancy in Fig. 6, where the cluster-conditional distributions $p_k^r(E)$ in (7) of the $K = 7$ partition with $\Delta \tau = 0$ are compared with the distributions of the corresponding partition with $\Delta \tau = 1000$ days. The latter, which are optimal for energy (see Fig. 5), have significantly smaller spread about their means, and as a result carry more information beyond climatology than their $\Delta \tau = 0$ counterparts (see also Table 2). Nevertheless, at least some of the regimes in the $\Delta \tau = 0$ model (namely, states 1, 4, and 7) have reasonably well-defined energy.

The key objects that facilitate our study of the physical properties of the regimes are the cluster-conditional mean and standard deviation of the streamfunction anomaly, $h_k^r(r) = \langle h^r(r, t) \rangle_k$ and $\sigma_k(r) = \langle (h^r(r, t) - h_k^r(r))^2 \rangle_k^{1/2}$, which are shown in Figs. 7 and 8 respectively for $K = 3$ and $K = 7$. Here, $\langle \cdot \rangle_k$ denotes expectation value with respect to $P_k^r$ (7) at $\tau = 0$, which, by ergodicity (3), can be evaluated by taking temporal averages conditioned on $S(t) = k$. Moreover, to examine the relationship between the circulation patterns in the regimes to the EOFs, it instructive to consider the projection coefficients $\chi^k_i = \int dr\text{ EOF}(r)_i h_k^r(r)$ in the basis of orthonormal EOFs, which are displayed in Fig. 9.

First, it is clear from Fig. 7 that the circulation regimes identified by the running-average partitioning algorithm with $K = 3$ are in good agreement with the semi-empirical phenomenology established for 1.5-layer double-gyre ocean models (McCalpin and Haidvogel 1996; Berloff and McWilliams 1999). Specifically, state 1, which has a low expected value of energy, $E_1 = \langle E(t) \rangle_1 = 3.5 \times 10^{17}$ J, features a meandering jet pattern; state 2, with $E_2 = 3.9 \times 10^{17}$ J resembles the climatological state $H(r)$ (see the corresponding $\chi^k_i$ coefficients in Fig. 9); and state 3 is dominated by a strongly, deeply-penetrating jet, and has large mean energy $E_3 = 4.2 \times 10^{17}$ J. As one might expect from the corresponding relative increase in information content (see Fig. 5), the basic spatial features of the $K = 3$ regimes are captured with significantly higher fidelity by the $K = 7$ model in Fig. 8. Here, the meanders in the low-energy state, $h_1^r(r)$, are sufficiently strong to prevent the jet from having a well-defined separation point in the western meridional boundary. The high-energy state, $h_3^r(r)$, resembles in the mean the extensional phase $h_3^r(r)$ of the $K = 3$ model, but the standard deviation of the $K = 7$ model, $\sigma_2(r)$, is concentrated more strongly around the $y = 0$ axis of the domain.

More quantitatively, the regimes of the $K = 7$ models can be grouped in categories according to the $\chi^k_i$ projection coefficients, plotted in Fig. 9. Ordering regimes in order of increasing mean energy, the following phenomenology emerges:

- **State $h_1^r$.** In this regime, the mean values of $PC_1$ and $PC_2$ are both small compared to their typical values in the remaining regimes. Therefore, the eddies associated with $PC_3$ and $PC_4$ distort the eastward jet significantly, resulting in a meandering jet configuration.

- **States $h_3^r$ and $h_4^r$.** These states both project strongly to EOF$_2$, while at the same time their projection coefficients to EOF$_1$ are small. Regimes of this type are not part of the $K = 3$ models, and feature a pronounced southward dip of the jet associated with EOF$_2$ (see Fig. 1). State $h_4^r$ differs from $h_3^r$ in that it has a large projection coefficient to EOF$_4$, which here interacts constructively with EOF$_2$, leading to a larger jet penetration length than in state $h_3^r$.

- **State $h_5^r$.** This regime is the analog to the mean-flow resembling state of the $K = 3$ model, in that its projection coefficients correlate strongly with those of the climatological average $H(r)$.

- **States $h_6^r$, $h_6^r$ and $h_7^r$.** These states all have large projection coefficients to EOF$_1$ ($\chi^1_i \gtrsim 0.5$), and as a result feature extensional jet configurations. The highest-energy state, $h_7^r$, is nearly identical to the corresponding state of the $K = 3$ model, at least in the mean. Regimes $h_6^r$ and $h_7^r$ feature progressively smaller jet-penetration lengths (due to interactions with EOF$_2$ and EOF$_3$), but in both cases the jet follows an approximately straight path as it travels eastward into the basin.

Turning to the temporal aspects of the switching process between the regimes, Fig. 10 illustrates that the cluster-affiliation sequence $S(t)$ (23) of the $K = 3$ partition of observation space with $\Delta \tau = 1000$ days leads to a natural splitting of the energy time series into persistent regimes (with decadal mean lifetimes; see Table 2), as expected from the high information content of that partition about energy. As remarked in §3a, imposing temporal regularity in $S(t)$ is frequently a challenge in AOS applications (e.g., standard $K$-means analysis of this data set results in unphysical, high-frequency transitions between the regimes), but it emerges here automatically by virtue of coarse-graining the training data set and the interval $\Delta \tau$ for initial cluster affiliation. It is important to emphasize, however, that persistence is not synonymous with skill. For instance, the $\delta_0$ score for $PC_1$ in Fig. 5 is a decreasing function of $\Delta \tau$, even though the persistence of the regimes exhibits a corresponding increase (as indicated by the drop in the number of transitions with $\Delta \tau$). Information theory allows one to tell when a persistent cluster-affiliation sequence actually carries information for prediction (or classification, as is the case for the $\tau = 0$ examples considered here), or is too crude of a description of the intrinsic low-frequency dynamics.
Figure 10 also shows the PC and energy affiliation sequences for the corresponding optimal $K = 7$ partitions, which, as stated above, have $\Delta \tau = 0$ (i.e., no coarse graining for initial cluster affiliation) and $\Delta \tau = 1000$ days, respectively. Here, two important observations are that (i) the persistence in $S(t)$, as measured by the mean time interval between transitions $A_k$ in Table 2, has a non-trivial dependence on $\Delta \tau$; and (ii) as stated in §4c, maximally informative regimes in the energy observable do not necessarily coincide in time with maximally informative regimes of the leading PCs encoding the large-scale circulation patterns. In particular, the discrepancy between energy and circulation regimes can be seen by comparing the portion of the $E(t)$ and $PC_i(t)$ time series in Fig. 10 with $t \in [200, 220]$ years.

5. Conclusions

In this paper we have developed a framework based on empirical information theory to assess ensemble predictive skill in long-range regime forecasting (§2), developed a running-average scheme for coarse-grained partitioning of observation spaces of ergodic dynamical systems (§3), and applied these tools to study the predictability of circulation regimes in an equivalent barotropic double-gyre ocean model (§4). In this framework, regimes are associated with coarse-grained partitions of observation spaces of atmosphere-ocean models, and long-range ensemble predictive skill is measured, via relative entropy, by the additional information beyond climatology about future values of observables given present knowledge of affiliation to a cluster in the coarse-grained partition.

Here, the time-dependent probability distributions for ensemble forecasts are conditioned only to cluster affiliation, and can be computed feasibly from the output of comprehensive GCMs with no dynamical model error beyond that of the GCM. Thus, the problem of specifying detailed initial conditions, which is frequently a limiting factor in long-range forecasts (Keenlyside et al. 2008; Hurrell et al. 2009; Meehl et al. 2009; Solomon et al. 2009), is avoided. In Paper II of this series (Giannakis and Majda 2011), the same probability distributions are used in related information-theoretic metrics to quantify the model error in reduced models for the transitions between the regimes.

The information-theoretic framework for predictive skill developed here can be applied irrespective of the algorithm used to partition the space of observations. As a concrete example, in this paper we have presented a partitioning algorithm where smoothing via running averages plays a central role in two distinct ways: Coarse-graining the training data used to compute the cluster coordinates in observation space, and averaging over an initial time interval $\Delta \tau$ the history of observations used to determine initial cluster affiliation, i.e., the coarse-grained initial conditions. This procedure, which shares certain aspects in common with kernel density-estimation methods (Silverman 1986), results in a multi-time family of partitions of observation space, from which optimal partitions for prediction are selected according to the information conveyed about the observable to be predicted and the prediction lead time.

Applying this toolkit to simulations of an equivalent barotropic, double-gyre ocean model (McCalpin and Haidvogel 1996; Berloff and McWilliams 1999) we find that

i. There exists significant skill in predicting certain key observables beyond the longest autocorrelation time of the state vector in observation space. In particular, the prediction PDFs for the energy and the leading PCs associated with the large-scale spatial structures of the eastward jet contain significant information beyond climatology for forecast lead times of order seven years (Fig. 4).

ii. The predictive information content in a coarse-grained partition depends strongly on the observable to be predicted and on the time interval used to determine the initial cluster affiliation. Specifically, the most skillful partitions for energy are obtained by assigning initial cluster affiliation based on the history of the system averaged over an interval $\Delta \tau$ spanning 1000 days, but the corresponding interval for the leading PCs is zero.

iii. Besides the three established circulation regimes in this class of models, which are characterized by meandering, mean-flow resembling, and extension phases of the jet, significant information content exists in a regime where the jet configuration is dominated by the second EOF, and in regimes associated with interactions of the jet with mid-basin eddies (Fig. 8). The information content of these models with $K > 3$ clusters remains significant for prediction lead times comparable to the longest decorrelation time of the PCs, $\tau_1 \simeq 1000$ days, but (as follows from the convergence of the $\delta^*_K$ curves for $K = 3$ and $K = 7$ in Fig. 4) the additional regimes contribute negligible skill beyond $K = 3$ models for longer forecasts.

A further desirable aspect of our approach is that it induces naturally a notion of temporal regularity in the affiliation sequence to clusters (Fig. 10), which is an issue frequently not taken into consideration in cluster analysis of AOS data (Christiansen 2007). Here, the smoothing of the system trajectory in observation space by running averaging results in a decrease in the number of transitions. In this regard, running averaging has a similar effect to a class of recently developed clustering schemes based on finite element (FEM) discretization and bounded-variation (BV) regularization (Horencz 2010).
In the future we plan to extend the analysis presented here for the equivalent barotropic model to comprehensive ocean models featuring baroclinic instabilities.

Acknowledgments.

This research of Andrew Majda is partially supported by NSF grant DMS-0456713, from ONR DRI grants N25-74200-F6607 and N00014-10-1-0554, and from DARPA grants N00014-07-10750 and N00014-08-1-1080. Dimitrios Gianannakis is supported as a postdoctoral fellow through the last three agencies. The authors wish to thank Rafail Abramov for providing simulation data for the 1.5-layer model, Illia Horenko and his group for stimulating conversations, and Paul Fischer for providing computational resources at Argonne National Laboratory. Much of this research was developed while the authors were participants in the long program at the Institute for Pure and Applied Mathematics (IPAM) on Hierarchies for Climate Science, which is supported by NSF, and in a recent month-long visit to the University of Lugano. We thank four anonymous reviewers for constructive comments on an earlier paper which was restructured as the present series.

REFERENCES


Table 1. Parameters of the 1.5-layer ocean model, taken from McCalpin and Haidvogel (1996).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interfacial friction</td>
<td>$r = 10^{-7} \text{ s}^{-1}$</td>
</tr>
<tr>
<td>Hyperviscosity</td>
<td>$A_{ib} = 8 \times 10^{10} \text{ m}^4 \text{ s}^{-1}$</td>
</tr>
<tr>
<td>Coriolis parameters,</td>
<td></td>
</tr>
<tr>
<td>$f_0$</td>
<td>$7.27 \times 10^{-5} \text{ s}^{-1}$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$1.98 \times 10^{-11} \text{ m}^{-1} \text{ s}^{-1}$</td>
</tr>
<tr>
<td>Mean seawater density</td>
<td>$\rho_0 = 1027 \text{ kg m}^{-3}$</td>
</tr>
<tr>
<td>Reduced gravity</td>
<td>$g' = g \Delta \rho / \rho_0 = 0.02 \text{ m s}^{-2}$</td>
</tr>
<tr>
<td>Inverse Rossby radius</td>
<td>$\gamma = f_0 / (g' H^*)^{1/2} = (47,636 \text{ m})^{-1}$</td>
</tr>
<tr>
<td>Wind stress amplitude</td>
<td>$\tau_0 = 0.05 \text{ N m}^{-2}$</td>
</tr>
<tr>
<td>Wind asymmetry parameter</td>
<td>$\alpha = 0.05$</td>
</tr>
<tr>
<td>Domain size</td>
<td>$(L_x, L_y) = (3600, 2800) \text{ km}$</td>
</tr>
</tbody>
</table>


Fig. 1. The climatological mean, $H(r)$, its standard deviation, $\sigma(r)$, and the leading four streamfunction-metric EOFs, evaluated using an equilibrated realization of the 1.5-layer model (29) of length $T = 10,000$ years sampled every $\delta t = 20$ days. The contour levels in the panels for $H(r)$ and $\sigma(r)$ are spaced by 12.5 m, spanning the interval $[-150, 150]$ m. Contours are drawn every 12.5 arbitrary units in the panels for EOF$_i(r)$, which also indicate the corresponding eigenvalues and correlation times, respectively $\lambda_i$ and $\tau_i$ (see Fig. 3). Solid and dotted lines correspond to positive and negative contour levels, respectively. The separation point of the eastward jet is located near the coordinate origin, $r = (x, y) = (0, 0)$. 
Fig. 2. Empirical autocorrelation functions, $\rho_i(t) = \int_0^T dt' PC_i(t')PC_i(t' + t)/T$, and equilibrium PDFs, $p_{eq}(PC_i)$, of the leading four streamfunction PCs. Among these PCs, only $PC_3$ has significantly negative values in $\rho_i(t)$. All the autocorrelation functions of $PC_i$ with $i \in [5,20]$ (not shown here) take negative values. Note that $p_{eq}(PC_i)$ are all unimodal, yet the system exhibits long-lived affiliations to regimes (see Fig. 10).

Fig. 3. Eigenvalues, $\lambda_i$, and correlation times, $\tau_i$, of the leading 20 EOFs/PCs of the streamfunction. The eigenvalues and EOFs are the solutions of the eigenproblem $\int dr' C(r,r')\text{EOF}_i(r') = \lambda_i \text{EOF}_i(r)$ associated with the covariance matrix $C(r,r') = \int_0^T dt h'(r,t)h'(r',t)/T$, where $h'(r,t)$ is the streamfunction anomaly (Holmes et al. 1996; Berloff and McWilliams 1999; Abramov et al. 2005). With this definition, the physical dimension of the $\lambda_i$ is (length)$^2$. The correlation times are given by $\tau_i = \int_0^T dt \rho_i(t)$, where $\rho_i$ is the autocorrelation function of the corresponding PC. Here, the eigenvalue and correlation-time plots are scaled respectively by the values of the leading EOF/PC, viz., $\lambda_1 = 28.96$ m$^2$ and $\tau_1 = 1165$ days.
Fig. 4. The information content (predictive skill score) $\delta_\tau$ (13) in $K = 3$ and $K = 7$ partitions of observation space (4) as a function of prediction horizon $\tau$ for the energy $E$ and the leading two PCs. Two values for the running-average interval for initial cluster affiliation are displayed ($\Delta\tau = 0$ and 1000 days), as well as the optimal skill score $\delta^*_\tau$ for various values of $\Delta\tau$ in the interval $[0, 1000]$ days. In all cases, the running-average interval for coarse-graining the training data set is $\Delta t = 1000$ days. The $\delta_\tau$ curves for energy with $\Delta\tau = 0$ illustrate that the decay of relative entropy to equilibrium may be non-monotonic; a behavior that cannot be replicated by Markov models (see Fig. 2 of Paper II). The $K = 7$ partitions have higher information content than the $K = 3$ ones in the leading PCs (i.e., the large-scale structures in the flow) for $\tau \lesssim 600$ days, or about half the decorrelation time of the leading PC (see Fig. 3). However, $K = 7$ contributes essentially no additional skill beyond $K = 3$ for decadal forecasts.
Fig. 5. The dependence of the number of transitions and the relative-entropy skill score $\delta_\tau$ (13) on the running-average interval $\Delta \tau$ (initial data for prediction), evaluated at time $\tau = 0$ for the energy $E$ and leading four PCs for models with $K \in \{3, 6, 7, 8\}$ clusters. The running-average interval for coarse-graining the training data is $\Delta t = 1000$ days and 200 days, respectively in the top and bottom set of panels.
Energy distribution in the regimes

![Energy distribution graph](image)

**Fig. 6.** Conditional PDFs $p_k^h(E)$ of the energy for $K = 7$ partitions with running-average interval for initial cluster affiliation $\Delta t = 1000$ days (thick solid lines) and $\Delta t = 0$ (dotted lines). In each panel, a PDF from the $\Delta t = 1000$ days partition and a PDF from the $\Delta t = 0$ partition are shown together, following the ordering convention in Table 2 (i.e., state 3 of the $\Delta t = 0$ partition is paired with state 2 of the $\Delta t = 1000$ days partition, and vice versa). The climatological distribution, $p_\text{eq}(E)$ is also plotted in thin solid lines for reference. These partitions are respectively optimized for energy ($\Delta t = 1000$ days) and the leading PCs ($\Delta t = 0$); see Fig. 5 and Table 2. In both cases, the running-average interval in the training stage is $\Delta t = 1000$ days.

Circulation regimes for $K = 3$, clustering optimized for energy

![Circulation regimes graph](image)

**Fig. 7.** Mean streamfunction anomaly, $h'_k(r)$, and its standard deviation, $\sigma_k(r)$, conditioned on the clusters of the $K = 3$ partition in Table 2. The contour-level spacing for $h'_k(r)$ and $\sigma_k(r)$ is 25 m and 10 m, respectively. Solid and dotted lines respectively represent positive and negative contour levels. This partition of observation space has been evaluated using running-average windows of duration $\Delta t = \Delta t = 1000$ days, and is optimized for maximal information content beyond climatology about energy (see Fig. 5). The spatial features of the circulation regimes identified here via running-average $K$-means clustering are in good agreement with the meandering ($h'_1$), mean-flow resembling ($h'_2$), and extensional ($h'_3$) phases of the jet in the McCalpin and Haidvogel (1996) phenomenology, with correspondingly low, moderate, and high values of mean energy (see Fig. 10).
Fig. 8. Mean streamfunction anomaly, $h'_k(r)$, and its standard deviation, $\sigma_k(r)$, conditioned on the clusters of the $K = 7$ partition of Table 2 with $\Delta \tau = 0$ (i.e., no temporal coarse graining for initial cluster affiliation). This partition of observation space is optimized for information content about the leading PCs (see Fig. 5). The contour-level spacing for $h'_k(r)$ and $\sigma_k(r)$ is 25 m and 10 m, respectively. Solid and dotted lines respectively represent positive and negative contour levels. States $h'_1$, $h'_5$, and $h'_7$ of this model are the analogs of the meandering, mean-flow resembling and extensional regimes in the $K = 3$ model in Fig. 7. The spatial configuration of the jet in state $h'_2$ is dominated by the second EOF (see Fig. 1), and has no counterpart in the $K = 3$ model. States $h'_2$, $h'_4$, and $h'_6$ respectively feature interactions of the jets in states $h'_5$, $h'_3$, $h'_7$ with mid-basin eddies in EOFs 3 and 4 (see Fig. 9).
Projection of the circulation regimes onto the EOF basis

\[ \chi_{ki} \]

\[ K = 7, \text{ state } 1 \]
\[ K = 3, \text{ state } 1 \]
\[ K = 7, \text{ state } 5 \]
\[ K = 3, \text{ state } 3 \]
\[ K = 7, \text{ state } 7 \]
\[ K = 7, \text{ state } 3 \]

EOF index \[ i \]

1 2 3 4 5 6

climatology

Fig. 9. Projection coefficients \( \chi_{ki}^k \) of the cluster-conditional mean streamfunction anomaly \( h_1^k(r) \) onto the leading six unit-norm EOFs (i.e., \( \int dr \text{EOF}_i^2(r) = 1 \)) for the \( K = 3 \) model in Fig. 7 (dashed lines, \( \Delta \tau = 1000 \) days) and the \( K = 7 \) model in Fig. 8 (solid lines, \( \Delta \tau = 0 \)). The projection coefficients of the climatological mean state \( H(r) \) are plotted in the lower-right panel for reference. In the phenomenology of McCalpin and Haidvogel (1996) state 2 of the \( K = 3 \) partition and state 5 of the \( K = 7 \) partition would be referred to as intermediate-energy or mean-flow resembling states due to the high correlation of their \( \chi_{ki}^k \) coefficients with those of \( H(r) \).
Fig. 10. Time series of the energy $E$ and the leading two PCs spanning a 200-year interval. The thick horizontal lines show the discrete energy and PC affiliation sequences, $\langle E \rangle_{S(t)}$ and $\langle PC_i \rangle_{S(t)}$, where $\langle \cdot \rangle_k$ denotes cluster-conditional expectation value, and $S(t)$ is the cluster-affiliation sequence in (23). Throughout, the running-average window in the training stage is $\Delta t = 1000$ days, and regimes are ordered in order of increasing $\langle E \rangle_k$. In the $K = 7$ panels, the running-average window $\Delta \tau$ is chosen so that the corresponding partitions of observation space contain high information about energy ($\Delta \tau = 1000$ days), or the leading PCs ($\Delta \tau = 0$; see Fig. 5 and Table 2).
Table 2. Properties of representative three-state and seven-state partitions of the observation space of the 1.5-layer model spanned by the leading 20 PCs (17) of the streamfunction, optimized for high information content about the energy $E_k$ or the leading PCs (see Fig. 5). In each case, listed are the equilibrium probability for cluster affiliation, $\pi_k$ (9), the mean regime lifetime, $\Lambda_k$, the cluster-conditional expectation value of the energy, $E_k$, and the relative-entropy skill metric, $D_k^0$ (8) evaluated at time $\tau = 0$ for $E$ and the leading two PCs. The weighted means by $\pi_k$ (i.e., the superensemble means) of $E_k$, $\Lambda_k$, and $D_k^0$, as well as the normalized skill scores $\delta_0$ from (13), are shown in the right-hand column. The highest scores for each variable are typeset in boldface font. Throughout, the clusters are indexed in order of increasing expected energy, $E_k$, and are arranged column-wise so that the clusters in a given column have similar projection coefficients $\chi_i^k$ of the streamfunction in the EOF basis (see Fig. 9); e.g., clusters (1, 2, 3) of the $K = 7$ models correspond to clusters (1, 2, 3) of the $K = 3$ model. Because some of the clusters of the $K = 7$ models are close in $E_k$, the grouping of the regimes based on $\chi_i^k$ does not preserve their ordering with respect to $E_k$.

\begin{tabular}{|c|c|c|c|}
\hline
$K = 3$, $(\Delta t, \Delta \tau) = (1000, 1000)$ days (partition optimized for $E$) & & & \\
State & 1 & 2 & 3 & Mean \\
$\pi_k$ & 0.222 & 0.607 & 0.171 & \\
$\Lambda_k$/days & 2080 & 3320 & 2300 & 2870 \\
$E_k/10^{17}$ J & 3.510 & 3.906 & 4.211 & 3.870 \\
$D_0^0(E)$ & 1.220 & 0.311 & 1.210 & 0.667 \\
$\delta_0(E)$ & 0.736 & 0.263 & 0.409 & \\
$D_0^0(PC_1)$ & 0.306 & 0.063 & 0.918 & 0.123 \\
$\delta_0(PC_1)$ & 0.319 & 0.218 & & \\
$D_0^0(PC_2)$ & 0.298 & 0.061 & 0.902 & \\
$\delta_0(PC_2)$ & 0.337 & & & \\
\hline
$K = 7$, $(\Delta t, \Delta \tau) = (1000, 0)$ days (partition optimized for PCs) & & & \\
State & 1 & 2 & 3 & 4 & 5 & 6 & 7 & Mean \\
$\pi_k$ & 0.112 & 0.194 & 0.140 & 0.159 & 0.171 & 0.117 & 0.108 & \\
$\Lambda_k$/days & 334 & 200 & 172 & 193 & 192 & 191 & 374 & 226 \\
$D_0^0(E)$ & 1.070 & 0.318 & 0.294 & 0.364 & 0.285 & 0.371 & 1.080 & 0.489 \\
$\delta_0(E)$ & 0.624 & 0.583 & 0.688 & & & & & \\
$D_0^0(PC_1)$ & 0.666 & 0.395 & 0.750 & 0.406 & 0.328 & 0.677 & 1.180 & 0.530 \\
$\delta_0(PC_1)$ & 0.663 & 0.515 & 0.688 & & & & & \\
$D_0^0(PC_2)$ & 0.726 & 0.230 & 0.146 & 0.462 & 0.548 & 0.176 & 0.220 & 0.357 \\
$\delta_0(PC_2)$ & 0.510 & 0.515 & 0.515 & & & & & \\
\hline
$K = 7$, $(\Delta t, \Delta \tau) = (1000, 1000)$ days (partition optimized for $E$) & & & \\
State & 1 & 2 & 3 & 4 & 5 & 6 & 7 & Mean \\
$\pi_k$ & 0.145 & 0.203 & 0.120 & 0.161 & 0.177 & 0.109 & 0.086 & \\
$\Lambda_k$/days & 1220 & 802 & 619 & 798 & 916 & 1030 & 1650 & 957 \\
$D_0^0(E)$ & 1.4589 & 0.628 & 0.891 & 0.549 & 0.790 & 0.495 & 2.018 & 0.899 \\
$\delta_0(E)$ & 0.834 & 0.362 & 0.515 & & & & & \\
$D_0^0(PC_1)$ & 0.287 & 0.150 & 0.419 & 0.281 & 0.214 & 0.249 & 1.510 & 0.352 \\
$\delta_0(PC_1)$ & 0.352 & 0.186 & 0.311 & & & & & \\
$D_0^0(PC_2)$ & 0.352 & 0.024 & 0.182 & 0.045 & 0.358 & 0.069 & 0.352 & 0.186 \\
$\delta_0(PC_2)$ & 0.324 & 0.311 & 0.311 & & & & & \\
\hline
\end{tabular}