Computational Geosciences manuscript No.
(will be inserted by the editor)

Analyzing state dependent model-data comparison in multi-regime systems

Alfredo L. Aretxabaleta · Keston W. Smith

Received: date / Accepted: date

Abstract An approach to analyze regime change in spatial time series data sets is followed and extended to jointly analyze a dynamical model depicting regime shift and observational data informing the same process. We analyze changes in the joint model-data regime and covariability within each regime. The method is applied to two observational data sets of equatorial sea surface temperature (TAO/TRITON array and satellite) and compared with the predicted data by the ECCO-JPL modeling system.

Keywords Skill assessment · Data clustering · Gaussian Mixture Models · ENSO

1 Introduction

The size and complexity of observational data sets are increasing constantly. Along with observations, we have ever more spatially resolved dynamical models of processes measured in spatial data sets. The best strategy for confronting physical models with data and the purpose of the comparison of models versus data remain as developing questions. Beyond simply obtaining a misfit, likelihood, or some other gross evaluation of the credibility of the model solution, we desire to know where, when, and why a model is performing poorly. While this is a simple idea, it is often explained with snapshots or a detailed analysis of an arbitrary episode because the full time series is too large and complex to analyze in its entirety.

Several methods have been proposed to analyze both stationary and non-stationary time series (e.g., [1]). Traditionally analysis of spatial and temporal patterns in geophysical spatial time series is carried out with Empirical Orthogonal Functions (EOFs, theory in [2] and examples of applications in [3,4]). An EOF analysis provides the leading eigenvectors of the temporal covariability of the data and then interprets these EOFs as the response to known physical processes. The eigenvectors are the “modes” of variability and their temporal “amplitude” functions define the temporal structure

Alfredo L. Aretxabaleta
Instituto de Ciencias del Mar - CSIC, Barcelona, Spain E-mail: alfredo@icm.csic.es

and Keston W. Smith
Woods Hole Oceanographic Institution, Woods Hole, MA, USA

of variability. In many studies [e.g., [5–7]] authors compare EOFs obtained from data to EOFs from models, and use their agreement as evidence of the fidelity of the model with respect to important physical processes. There are shortcomings to this approach. Firstly, the model may produce the correct modes, but at the wrong times because of phase errors in the model. Secondly, EOFs are an analysis of covariance and as such they do not consider the non-Gaussian properties of the spatial distribution. In the case of using the EOF method for non-Gaussian distributions, it provides an analysis of the best Gaussian approximation to the distribution.

In this study, we present a method, the Joint Empirical Orthogonal GAussian Mixture Model Analysis (JEO-GAMMA), for analyzing the joint distribution of spatial time series of model predictions and data. The outcome is a set of easy to interpret representations showing the modes of spatial covariability in the model and data. The method accounts for non-Gaussian state distributions, or regime change, by analyzing variability about a small number of mean states. In a previous related study [8], Expectation Maximization (EM) was used to estimate the parameters of a Gaussian Mixture Model providing a distinct temporal decomposition relative to EOF analysis. We showed that while conventional EOF analysis was ambiguous for regime separation, EM produced clear separation of the spatial modes facilitating the physical interpretation of the data.

The remainder of the paper is organized as follows: In Section 2, we define the mathematical structure Gaussian Mixture Models (GMM) and describe the approach to fit GMM to data sets. In Section 3, we apply the JEO-GAMMA method to a combination of data from equatorial Pacific sea surface temperature (SST) from the TAO/TRITON array and a global circulation model describing the same region. In Section 4, we extend the method to a higher dimensional dataset of the same region using satellite SST and an expanded model solution. Conclusions and possible extensions of the method are given in Section 5.

2 Methods

2.1 Gaussian Mixture Models

A Gaussian Mixture Model is a probabilistic model for which the probability density function is a combination of two or more Gaussian distributions. Let D denote a discrete spatial time series of observations with time in the columns and some set of fixed positions in the rows. Let M denote a model whose time and space domain covers the region of D. In general, the relationship between D and M is given by $D = H(M) + \delta$, where δ is the difference between model and data, and H is a nonlinear measurement operator. In our case, we assume D and M are spatially and temporally collocated (i.e., H is the identity matrix). We augment the matrix of data with the model's approximation to the data,

$$\psi = [D \;\; M] \tag{1}$$

Next we fit a mixture model to the joint data-prediction data set, ψ. For an n_c component Gaussian mixture model, we have in general

$$p(\psi|\mu^1, ...\mu^{n_c}, \Sigma^1, ...\Sigma^{n_c}\tau^1, ..\tau^{n_c}) = \sum_{k=1}^{n_c} \tau^k \frac{exp(-\frac{1}{2}(\psi - \mu^k)^T[\Sigma^k]^{-1}(\psi - \mu^k))}{\sqrt{(2\pi)^{2n_d}|\Sigma^k|}} \tag{2}$$

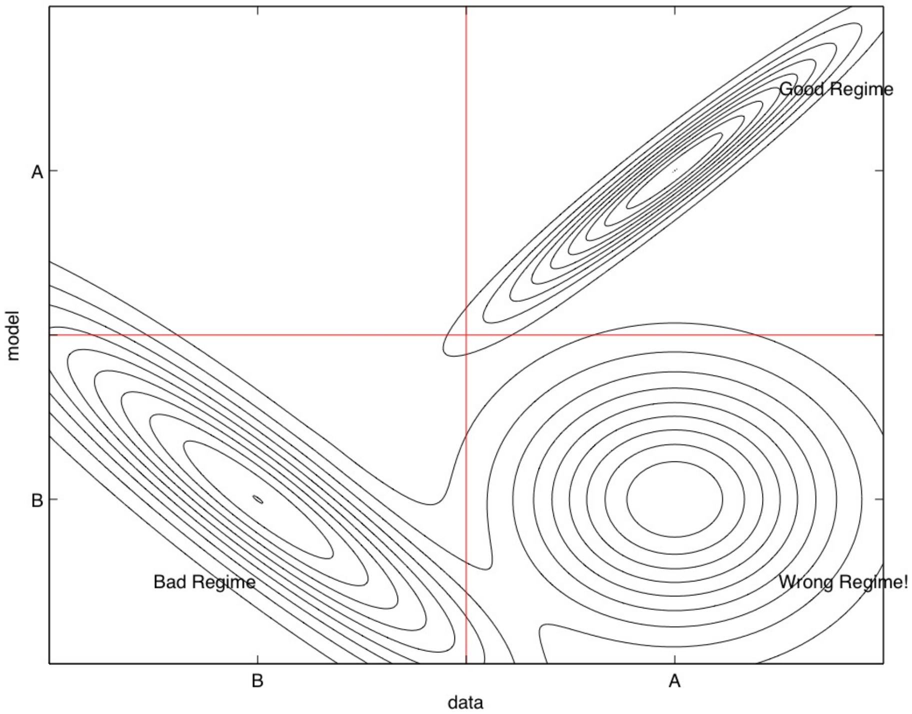

Fig. 1 Idealized depiction of joint model-data probability distribution. Here we show three possible model regimes, a "good" regime (upper right quadrant) in which the model and data are both in physical regime A, and the model and data are positively covarying. A "bad" regime (lower left quadrant) in which the model and data are both in regime B but the model and data are anti covarying. A bad regime can appear not only when model and data are anti correlated, but also when the model and data vary in different ways. Lastly we show a case of the Data being in regime A while the model is in regime B - "wrong regime" (lower right quadrant).

An underlying assumption is the stationarity of the distribution. For non-stationary cases, a trend parameter can be added to each regime mean or if there is a global trend, it can be extracted before the EM analysis. We use limited length time series for which the assumption of stationarity is appropriate. The use of this method for non-stationary time series goes beyond the scope of this study.

We use the Expectation-Maximization (EM) algorithm, outlined in [8] and Apendix A, to find the best GMM describing the joint distribution of the model and data. In previous studies, EM was used to estimate missing values for oceanographic datasets [9, 10]. In the present study, by using EM to estimate the parameters of the GMM, we are able to use EM to identify regimes in spatial time series and analyze the variability within each regime. After we have found the number of components, n_c, component distributions (mean and covariance), $G(\mu^k, \Sigma^k)$, and their respective likelihoods, τ^k, we can conduct the EOF analysis on the Σ^k and separate them into their data and model parts. n_d is the number of time series of length n_t.

The goal is to produce a comparison of the joint data-model distribution that characterizes the separation into the regimes observed in the combined matrix (Equation 1). In an optimal prediction, the "good" regime (Figure 1) will be predicted by the model and the statistical characteristics of the data during that regime will be appropriately reproduced by the model. A regime can be bad in several ways. Firstly, the model may have a strong bias within a particular regime. Secondly, the model may not covary with the data within a regime, either because the magnitude or direction of covariance represents an error in the model prediction. Finally, in an extreme case the "wrong" regime will be predicted by the model. A model that results in "wrong" regime estimates should not be used for non-linear applications that require proper characterization of different regimes. A model that exhibits deficiencies (bias, poor covariability) in its regime estimation may or not be useful depending on the application and the nature of the deficiencies.

2.2 Determining the number of regimes

A difficulty of the clustering approach is the lack of a generalized statistically principled method for determining the number of clusters. Several methodologies have been proposed to address this issue using empirical or data-based approaches.

A first option is the use of the empirical Akaike's Information Criterion (AIC, [11]). In general, $AIC(k) = 2D_k - 2log(\hat{p})$, where D_k is the number of free parameters in the statistical mixture model, and $\hat{p}$ is the maximized likelihood function for the estimated model. The goal is to rank several competing models according to their AIC, with the best being the one with the lowest AIC. The goodness of fit improves as the number of estimated free parameters (number of clusters) is increased. AIC aims at optimizing goodness of fit while including a penalty to discourage overfitting that increases with increasing number of clusters.

A second empirical approach is the Bayesian Information Criterion (BIC, [12]). The BIC approximates the total probability (Bayes factor) of a probability distribution under some set data,

$$BIC(k) = -2log(\hat{p}(\psi|\mu_1, ..., \mu_k, \Sigma_1, ..., \Sigma_k, \tau_1, ..., \tau_k)) - D_k log(n_t) \quad (3)$$

For the full mixture model with k components and n model-data time series, $D_k = k(n(n-1)/2 + n) + k - 1$, where $kn(n-1)/2$ of those are for the parameters of the covariance matrix, kn for the means of each distribution, and $k-1$ for the τ_j. The preceding "data" refers to the combination of model and data. As for the AIC, the model with the lower value of BIC is the one to be preferred. The penalty preventing overfitting is larger in the BIC than in the related AIC.

A completely different approach uses a data-driven method to estimate the number of clusters. In one example, [13] calculate the cross-validated likelihood. The method pre-analyzes the data to estimate a posterior probability distribution for the number of clusters. In cross-validation, the data is repeatedly divided into two subsets, one to fit the model and the other to estimate performance. The procedure is repeated multiple times and the results for each subsampling are averaged to obtain a mean estimate of the number of clusters. A second example of data-driven method [14] finds uncertainties on the estimated parameters to determine the number of regimes. It calculates confidence intervals of the mixing proportions based on order statistics by producing multiple estimates of the parameters. The main inconvenience of these two data-driven approaches is that the quality of the separation depends on the number of cross-validation subsamplings or uncertainty estimates. Both methods require at least

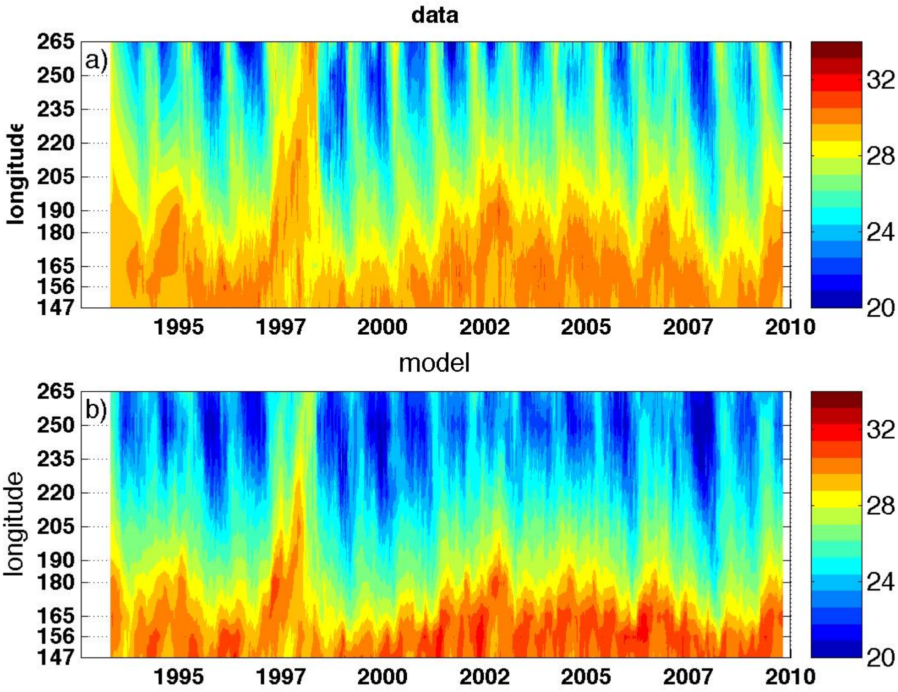

Fig. 2 Data (**a**) and ECCO model (**b**) SST (°C) for period of co-availability for each longitudinal station. The x-axis indicates years.

one hundred samples, which for high-dimensional problems as the ones presented in this study, will result in the approach being too computationally expensive.

In this study, we use the Bayesian Information Criterion to identify the number of component distributions in the data set because of its simplicity, reproducibility and relatively low computational cost. This approach has been shown to optimally estimate the quantity of clusters ([12, 15, 16]).

3 Application to Equatorial Pacific

3.1 Data and Model

A subsample of the TAO array data consisting of sea surface temperature (SST) from the equatorial Pacific moorings (including stations along the Equator, and at 2°N and 2°S) is used. Data from this array has been extensively used to understand the El Niño/Southern Oscillation (ENSO) dynamics [17, 18]. In this study the data (Figure 2a) is block averaged between 2°N and 2°S for each longitude resulting on a set of 611 temporal instances (to match model output) for each of the 10 longitudinal points considered.

The model is a non-assimilative global model solution provided by Estimating the Circulation and Climate of the Ocean (ECCO-JPL, [19,20]) which is based on the MIT general circulation model (MITgcm). The model has a horizontal resolution that varies between 0.3 and 1 degree. As with the data, we average the model solution between 2°S and 2°N. The time step of the time average model output fields is 10 days. The top layer of the model temperature (5 meters) is taken to be the best approximation to the observed SST. The period of co-occurrence with the TAO data stretches from spring of 1993 to fall of 2009 (Figure 2b). The model shows a tendency to be colder than the observations at the eastern stations and slightly warmer at the western ones.

3.2 Results

The BIC selects for three component distributions in the joint model-data distribution (the same number as in [8]). The three regimes show similar spatial patterns that are clearly present in the original data with warmer temperatures in the western stations (Figure 3). The spatial distribution of the means differs only slightly between data and model. We call the component most predominant in time Regime A and it is present 55% of the time. The second most frequent component (Regime B) is identified 34% of the time and the third component (Regime C) corresponds to the remaining 11%. Examining the time-varying probability (most often we find $w^k(t) = 0$ or $w^k(t) = 1$) of being in each regime (Figure 3,a5,b5,c5) and comparing them with the NOAA Multivariate ENSO Index (MEI, [21]), we can relate the different regimes to the different ENSO states. Positive (*negative*) MEI corresponds to El Niño (*La Niña*) conditions when it exceeds a certain threshold that in our representation is normalized to be 1 (*-1*) and otherwise corresponds to "normal conditions". Thus, the three regimes correspond to normal conditions (Regime A), La Niña (Regime B), and El Niño (Regime C). All the regimes means (Figure 3,a1,b1,c1) show a strong cold bias in the model solution (red line) east of the international date line that ranges $1 - 2$ °C.

The first mode of the EOF analysis of Regime A (associated with "normal conditions", Figure 3,a2) shows the predominant covariability is in the eastern stations. The model variability corresponds with the observed variability except for in the easternmost station. The model-data covariability is coherent across the entire spatial extension of the second EOF (Figure 3,a3). In the third EOF mode (Figure 3,a4), a component of variability in the easternmost station is not reproduced in the model resulting in model and data being anti-correlated. The model exhibits twice the observed variability in this mode east of 220 ($140°W$).

During La Niña conditions (Regime B) the model mean (Figure 3,b1) is slightly worse than during normal conditions reproducing the spatial structure but not the magnitude exhibiting a larger bias. Most of the variability associated with this regime is present in the first mode ($V = 14$) and the model-data discrepancies for this mode (Figure 3,b2) were similar in structure to the first mode of the Regime A. The model component of the covariability in the third mode of this regime (Figure 3,b4) differs significantly from the observed spatial structure by exhibiting a mode of model variability in the west not present in the data.

Finally, during El Niño (Regime C) the model displays the worse deficiencies. The model mean (Figure 3,c1, red line) resembles more the observed mean from Regime A than the mean for Regime C. All modes of variability include deterioration of the model

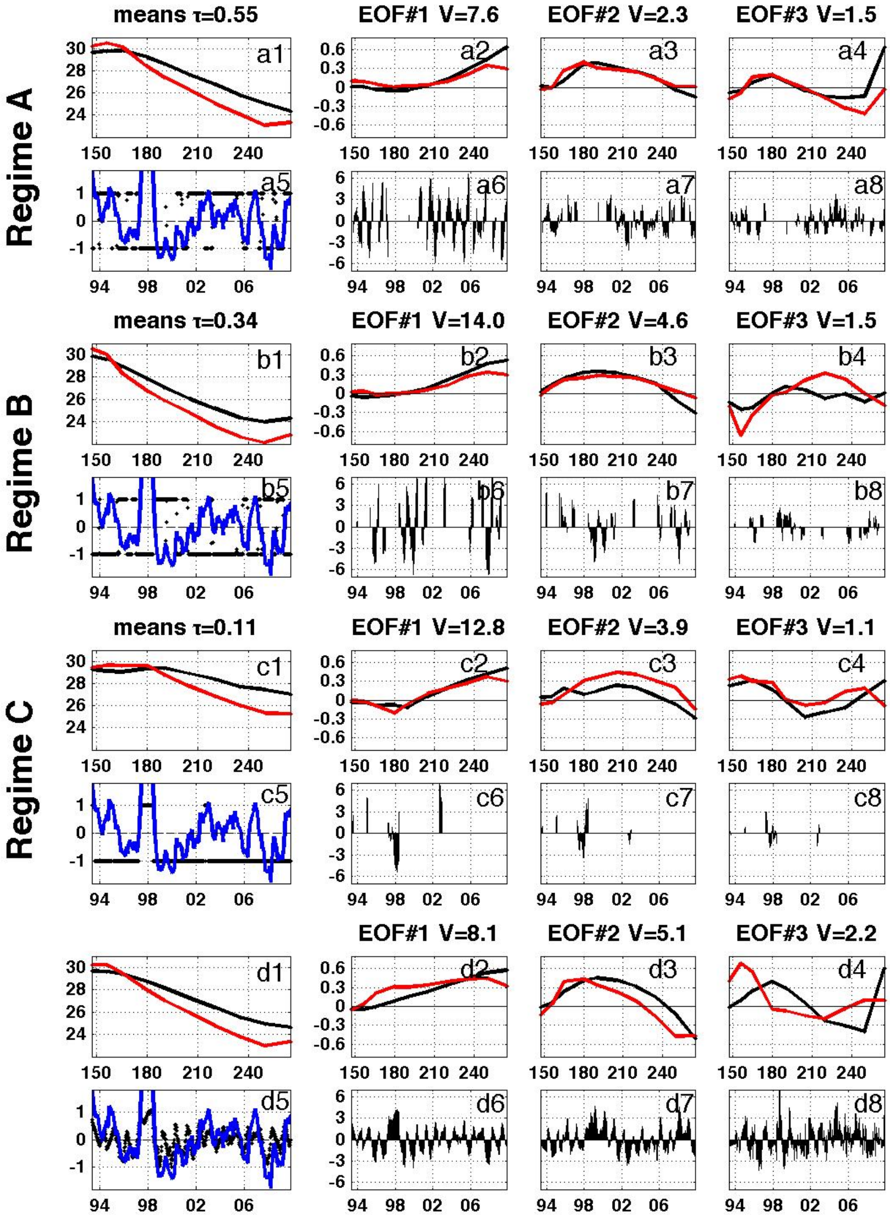

Fig. 3 The three identified regimes from top to bottom in frequency: Regime A (55% frequency) is shown in the top 8 panels, Regime B (34%) in the next 8 panels and Regime C (11%) in the following 8 panels. The last eight panels correspond to the conventional EOF analysis of the entire dataset (no regime separation) for comparison. The data is in black and the model in red. For each regime the panels are: **1:** The longitudinal distribution of the data and model mean; **2,3,4:** The spatial distribution of the 1^{st}, 2^{nd} and 3^{rd} EOFs of the joint co-variability (the size of each mode is included in the title of each panel); **5:** Adjusted probability of the regime (the axis has been stretched so that for regime k, $w^{j=k} = 1$ and $w^{j \neq k} = -1$) and time series of normalized ENSO MEI Index (blue line) where positive (*negative*) values larger (*smaller*) than 1 (*-1*) correspond with El Niño (*La Niña*) conditions; **6,7,8:** The time varying amplitudes of the first three EOFs (valid only for periods when the regime has been separated). **d5** includes the normalized time-varying magnitude of the 1^{st} EOF in black.

skill with the first mode (Figure 3,c2) having problems around the date line, and the second and third modes (Figure 3,c3,c4) being poorly captured in most stations.

The joint temporal variability of model and data represented in the lower panels of each of the regimes allows the interpretation of the temporal changes for each regime. For instance for Regime C, the first EOF time-vaying amplitude (Figure 3,c6) separates the large 1998 El Niño from other smaller El Niño periods (1993, 1994, 2003). The second EOF (Figure 3,c7) separates the variability associated with the initiation of El Niño from the one associated with its breakdown.

When the entire data set is analyzed without the use of EM for regime separation, the resulting averages (Figure 3,d1) are very similar to the Regime A (normal conditions) averages. Using the conventional EOF analysis in the entire dataset, the longitudinal distribution of variability for the different modes present some differences from the modes of each of the regimes. The first EOF (Figure 3,d2) exhibits increased model variability (compared to the modes obtained after EM) in the region between the dateline and 220 ($140°W$). This is caused by the changes from regime to regime, as it is not present in any of the first modes obtained by the EM separation. The second EOF of the entire data set (Figure 3,d3) exhibits a similar longitudinal structure to the second modes of each of the regimes, while the third EOF (Figure 3,d4) is completely different.

The EM method provides a more accurate regime separation than using a conventional EOF approach (no EM used). When compare EOF and EM to the MEI Index, the conventional EOF method estimates the correct regime 73% of the time (Figure 3,d5) while the EM algorithm correctly predicts the ENSO state 92% of the time. Furthermore, the clear modal separation achieved by the EM analysis facilitates the physical interpretation of the data.

4 Higher dimensional application

One of the main concerns of this methodology is the applicability to larger data sets such as realistic model outputs and satellite observations. We conduct an additional experiment to compare daily high-resolution blended SST ([22, 23]) and the ECCO-JPL model solution (Section 3.1) in the same area of the Equatorial Pacific but now extending from 5°N to 5°S. The original 0.25°-resolution SST data is averaged to match the 0.3° latitudinal and 1° longitudinal resolution of the model resulting in 4000 spatial points and 659 temporal instances.

In theory, the computational cost of using the EM algorithm to separate the components of the GMM could be expensive for high dimensional problems. In practice, the extraction of the EOFs is also computationally intensive for these problems and in fact in this application the EM algorithm is only six times more costly than the basic EOF analysis. Clearly, the combined cost is high but we believe the improved results and the ease of interpretation compensate for the increased cost.

The method separates three components in the extended model-data distribution (Figure 4). The regimes in this case are very similar to the ones extracted in Section 3. The most predominant component (Regime A) is present 52% of the time, while the second (Regime B) is identified 36% of the time, and Regime C corresponds to the remaining 12%. As in the previous case, Regime A is consistent with "normal conditions", Regime B with La Niña, and Regime C with El Niño. The probability of each regime exhibits a binary behavior, with $w^k(t) = 0$ or $w^k(t) = 1$ most of the time.The

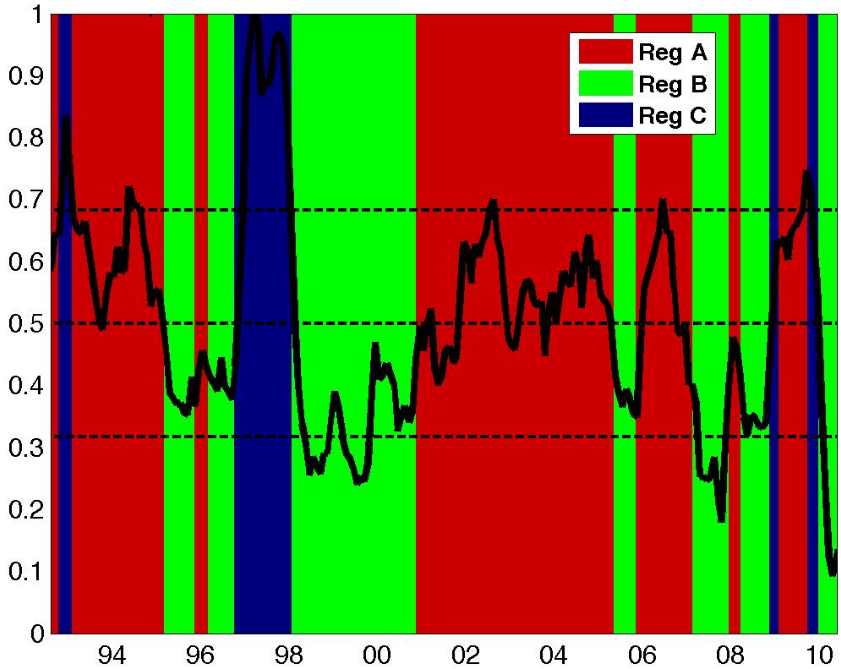

Fig. 4 Probability of the three separate regimes and time series of normalized ENSO MEI Index (black line) where values larger (*smaller*) than 0.66 (*0.33*) correspond with El Niño (*La Niña*) conditions. The x-axis indicates years.

EM algorithm is slightly worse than in Section 3 at predicting correctly the ENSO state (correct regime 84% of the time), because of the presence of additional variability associated with other processes such as the seasonal cycle.

The three regimes show different spatial patterns (Figure 5) with some common features present in both the satellite data and the model simulation: warmer temperatures in the west, slightly cooler temperatures in the southern than in the northern hemisphere. Regime A (normal conditions) exhibits a cold model bias (Figure 5c) in most of the domain with larger values along the Equator between 190-260 ($170 - 100°W$). The difference between the Regime A first EOF of data and model (Figure 5d,e) is significant with the model highest variability centered in a position to the northwest of the data and exhibiting a smaller maximum. In the case of Regime B (La Niña), the model bias (Figure 5h) is larger in magnitude but concentrated over a smaller area. The model first EOF of Regime B closely resembles the structure and magnitude of the data first EOF (Figure 5i,j). The model during El Niño (Regime C) exhibits a larger colder bias (Figure 5m) with its maximum concentrated around 260 ($100°W$). The model first EOF for Regime C (Figure 5o) exhibits the largest deficiencies failing to appropriately characterize its maximum in magnitude and longitudinal and latitudinal position. When the entire data set is analyzed (without regime partition), the model

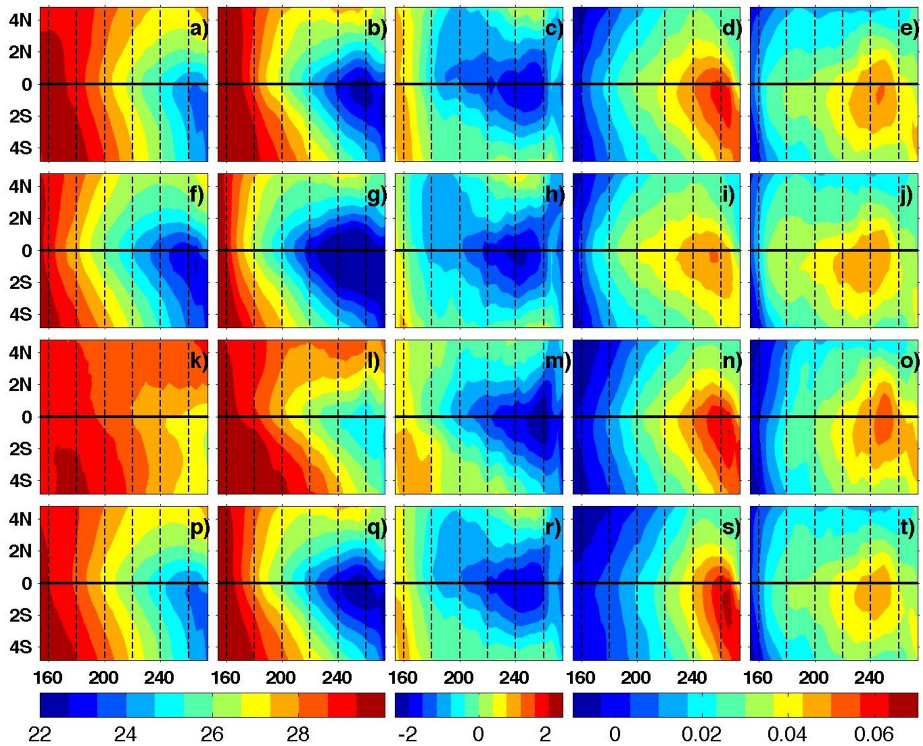

Fig. 5 Spatial distribution of the means, bias and 1^{st} EOF of the data and model solutions. The first, second and third rows corresponds to Regimes 1, 2 and 3, respectively. The fourth row corresponds to the entire dataset. The first column is the data mean; the second is the model mean; the third is the bias (model-data); the fourth is the 1^{st} EOF of the data; and the last column is the 1^{st} EOF of the model.

bias (Figure 5r) and the structure of the data and model first EOF (Figure 5s,t) closely mimic the results for Regime A.

In general, the model presents some deficiencies, especially during El Niño periods, that include sporadic poor correlation with the data and imperfect variability structure and magnitude representation. While these deficiencies can be severe in specific locations and times, the joint model-data distribution suggests the model is able to characterize the right regime for each of the three separate components.

5 Conclusions

As a generalization to EOF analysis, JEO-GAMMA allows for a non-Gaussian description of model-data joint distributions. The applications of the method extend from model skill estimations to improved regime separation.

The method allows the analysis of the variability in each component separately with an optimal and non-arbitrary procedure. The data-model comparison is therefore achieved inside the limits of the specific regime instead of having to concentrate in concrete periods or entire time series that include multiple regime signals. The separation of each regime permits the description of the predominant modes around clearly defined and statistically distinguishable means.

JEO-GAMMA can be summarized as a procedure to first objectively separate the different components (regimes) of a GMM using the EM methodology and then analyze the covariance in each regime using EOF analysis. Previous studies [13, 14] followed the reversed path, using EM to separate clusters inside EOF modes from geopotential height anomalies. We believe our approach is more appropriate for regime separation and skill assessment.

We demonstrate the applicability of the method for both small (TAO/TRITON vs ECCO-JPL model) and large (satellite SST vs model) data sets. The application of this methodology to extremely large datasets (millions of spatial datapoint) may require additional slight modifications by the implementation of high-dimensional data clustering algorithms (e.g., [24]). We believe these modifications to be small (if necessary) and therefore expect the method to be of great usefulness.

Therefore, the method represents an efficient and flexible approach for regime identification and analysis especially for model skill assessment. We believe that the preservation of the realistic multi-regime structure of a system should be encouraged in future statistical analysis of the ocean.

Acknowledgements The TAO/TRITON data were downloaded from the TAO Project Office server (http://www.pmel.noaa.gov/tao/). The satellite data were downloaded from the NCDC thredds server (http://nomads.ncdc.noaa.gov/thredds/catalog/oisst2/catalog.html). The state estimates were provided by the ECCO Consortium for Estimating the Circulation and Climate of the Ocean funded by the National Oceanographic Partnership Program (NOPP). This is a contribution to the SMOS Barcelona Expert Center on Radiometric Calibration and Ocean Salinity. Funding for this work was provided by Spanish National Program on Space, under contract ESP2005-06823-C05. A. Aretxabaleta has been additionally supported by a Juan de la Cierva grant of the Spanish Government. K. Smith was supported by NSF Grant DMS-0934653.

Appendices

A Expectation-Maximization

The EM algorithm is an iterative procedure to find the Maximum Likelihood Estimate of the parameters of a Gaussian Mixture Model by applying the following two steps:

Expectation step: The expected value for component k of the likelihood function, $w^k(t)$, is calculated under the current estimate of the parameters μ^k and Σ^k:

$$w^k(t) = \frac{e^{(-\frac{1}{2}(\psi-\mu^k)^T[\Sigma^k]^{-1}(\psi-\mu^k))}}{\sqrt{(2\pi)^{n_d}|\Sigma^k|}}, \tag{4}$$

$$w^k(t) \to \frac{w^k(t)}{\sum_j w^j(t)} \tag{5}$$

The $w^k(t)$ is used for the temporal description of the time series, being analogous to the temporal amplitudes produced by EOF analysis. In practice, we find that most often there is a tendency for binary behavior, with $w^k(t) = 0$ or $w^k(t) = 1$.

Maximization step: The optimal parameters that maximizes the current estimate given the data $\psi(t)$ is calculated. Note that τ^k, μ^k and Σ^k may be all maximized independently of each other since they appear in separate linear terms.:

$$\tau^k = \frac{n^k}{n_t} = \frac{\sum_t w^k(t)}{n_t} \tag{6}$$

$$\mu^k = \sum_t w^k(t)\psi(t)/n^k \tag{7}$$

$$\Sigma^k = \sum_t w^k(t)(\psi(t) - \mu^k)(\psi(t) - \mu^k)^T/(n^k - 1) \tag{8}$$

This procedure converges to a local maximum of the likelihood function [25]. The convergence to the global maximum is achieved by the repetition of the algorithm with random initial means. The mean of the first component is randomly chosen from the data points and the second and successive components are chosen such that their states are farthest from the precedent means.

References

1. Huang, N. E. , Shen, Z., Long, S. R., Wu, M. C., Shih, E. H., Zheng, Q., Tung, C. C., Liu, H. H., The Empirical Mode Decomposition Method and the Hilbert Spectrum for Non-stationary Time Series Analysis, Proc. Roy. Soc. London, A454, 903–995, (1998).
2. Emery, W. J., Thomson, R. E., Data Analysis Methods in Physical Oceanography, Pergamon, pp 634, (1997).
3. Bretherton, C. S., Smith, C., Wallace, J. M., An intercomparison of methods for finding coupled patterns in climate data, J. Climate, 5, 541–560, (1992).
4. Tourre, Y. M., White, W. B., ENSO signals in global upper-ocean temperature, J. Phys. Oceanogr., 25, 1317–1332, (1995).
5. Goswami, B., Shukla, J., Predictability of a coupled ocean-atmosphere model, J. Climate, 4, 3–22, (1991).
6. Murtugudde, R., Busalacchi, A., Interannual Variability of the Dynamics and Thermodynamics of the Tropical Indian Ocean, J. Climate, 12, 2300–2326, (1999).
7. Barnett, T. P., Comparison of near-surface air temperature variability in 11 coupled global climate models, J. Climate, 12, 511–518 (1999).
8. Smith, K. W., Aretxabaleta, A. L., Expectation-maximization analysis of spatial time series, Nonlin. Processes Geophys., 14, 73–77, (2007).
9. Houseago-Stokes, R. E., Challenor, P. G., Using PPCA to Estimate EOFs in the Presence of Missing Values, J. Atmos. Oceanic Technol., 21, 14711480, (2004).
10. Kondrashov, D., Ghil, M., Spatio-temporal filling of missing points in geophysical data sets, Nonlin. Processes Geophys., 13, 151159, (2006).
11. Akaike, H., A new look at statistical model identification, IEEE Transactions on Automatic Control, 19, 716–723, (1974).
12. Schwarz, G., Estimating the dimension of a model, Ann. Stat., 6, 461–464, (1978).
13. Smyth, P., Ide, K., Ghil, M., Multiple Regimes in Northern Hemisphere Height Fields via Mixture Model Clustering, J. Atmos. Sci., 56, 3704–3723, (1999).
14. Hannachi, A., Tropospheric Planetary Wave Dynamics and Mixture Modeling: Two Preferred Regimes and a Regime Shift, J. Atmos. Sci., 64, 3521–3541, (2007).
15. Hu, X. L., Xu, L., A comparative study of several cluster number selection criteria, In: Proc. of IDEAL03, Lecture Notes in Computer Science, LNCS 2690, Springer-Verlag, 195–202, (2003).
16. Raftery, A., Dean, N., Variable selectionfor model-based clustering, J. Amer. Statist. Assoc., 101 (473), 168–178, (2006).
17. McPhaden, M. J., Busalacchi, A. J., Donguy, R. C. J. R., Gage, K. S., Halpern, D., Ji, M., Julian, P., Meyers, G., Mitchum, G. T., Niiler, P. P., Picaut, J., Reynolds, R. W., Smith, N., Takeuchi, K., The Tropical Ocean-Global Atmosphere observing system: A decade of progress, J. Geophys. Res., 103, 14169–14240, (1998).
18. McPhaden, M. J., Genesis and evolution of the 1997-1998 El Niño, Science, 283, 950–954, (1999).
19. Wunsch, C., Heimbach, P., Practical global oceanic state estimation, Physica D., doi:10.1016/j.physd.2006.09.040, (2007).

20. Wunsch, C., Heimbach, P., Ponte, R. M., Fukumori, I., the ECCO-GODAE Consortium Members, The Global General Circulation of the Ocean Estimated by the ECCO-Consortium, Oceanography, 22, 88–103, (2009).
21. Wolter, K., Timlin, M. S., Measuring the strength of ENSO - how does 1997/98 rank?, Weather, 53, 315–324, (1998).
22. Reynolds, R. W., Rayner, N. A., Smith, T. M., Stokes, D. C., Wang, W., An improved in situ and satellite SST analysis for climate. J. Climate, 15, 1609–1625, (2002).
23. Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., Schlax, M. G., Daily High-Resolution-Blended Analyses for Sea Surface Temperature, J. Climate, 20, 5473–5496, (2007).
24. Bouveyron, C., Girard, S., Schmid, C., High-dimensional data clustering, Comput. Statist. Data Analysis, 52, 502–519 (2007).
25. Fraley, C., Raftery, A., Model based clustering, discriminant analysis, and density estimation, J. Amer. Statist. Assoc., 97 (458), 611–631, (2002).