Annales Geophysicae (2003) 21: 167–187 © European Geosciences Union 2003

Annales Geophysicae

Multivariate Empirical Orthogonal Function analysis of the upper thermocline structure of the Mediterranean Sea from observations and model simulations

S. Sparnocchia[1], **N. Pinardi**[2], **and E. Demirov**[3]

[1]Istituto Talassografico di Trieste-CNR, Viale R. Gessi 2, 34123 Trieste, Italy
[2]Corso di Scienze Ambientali, University of Bologna, Ravenna, Italy
[3]Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

Received: 1 October 2001 – Revised: 15 May 2002 – Accepted: 5 July 2002

Abstract. Multivariate vertical Empirical Orthogonal Functions (EOF) are calculated for the entire Mediterranean Sea both from observations and model simulations, in order to find the optimal number of vertical modes to represent the upper thermocline vertical structure. For the first time, we show that the large-scale Mediterranean thermohaline vertical structure can be represented by a limited number of vertical multivariate EOFs, and that the "optimal set" can be selected on the basis of general principles. In particular, the EOFs are calculated for the combined temperature and salinity statistics, dividing the Mediterranean Sea into 9 regions and grouping the data seasonally. The criterion used to establish whether a reduced set of EOFs is optimal is based on the analysis of the root mean square residual error between the original data and the profiles reconstructed by the reduced set of EOFs. It was found that the number of EOFs needed to capture the variability contained in the original data changes with geographical region and seasons. In particular, winter data require a smaller number of modes (4–8, depending on the region) than the other seasons (8–9 in summer). Moreover, western Mediterranean regions require more modes than the eastern Mediterranean ones, but this result may depend on the data scarcity in the latter regions.

The EOFs computed from the in situ data set are compared to those calculated using data obtained from a model simulation. The main results of this exercise are that the two groups of modes are not strictly comparable but their ability to reproduce observations is the same. Thus, they may be thought of as equivalent sets of basis functions, upon which to project the thermohaline variability of the basin.

Key words. Oceanography: general (water masses) – Oceanography: physical (hydrography; instruments and techniques)

Correspondence to: S. Sparnocchia
(sparnocchia@itt.ts.cnr.it)

1 Introduction

The Mediterranean Sea has been the site of major international scientific programs that have set the scene for the modern study of the thermohaline variability of this basin. A comprehensive historical in situ data set has been collected and disseminated (Medatlas, Fichaut et al., 1998) and numerical models have been extensively used to simulate the large-scale circulation at the seasonal and interannual time scales (Pinardi and Masetti, 2000). In addition, the Mediterranean Forecasting System Pilot Project (MFSPP, Pinardi et al., 2003) has begun investigations for the forecasting of the Mediterranean Sea large-scale circulation based upon the scientific knowledge of the processes, a near real-time observing system (Pinardi et al., 2002) and data assimilation scheme. This paper contributes to the description of the Mediterranean Sea thermohaline structure for general purposes and as part of the necessary knowledge required for the proper assimilation of in situ data with the Optimal Interpolation scheme used in MFSPP (Demirov et al., 2003).

The thermohaline vertical structure of world ocean basins is represented by evaluating pertinent temperature-salinity (T-S) diagrams obtainable from in situ data sets. Recent works (De Mey and Robinson, 1987; Fukumori and Wunsch, 1991) have shown that multivariate Empirical Orthogonal Functions (EOF) can efficiently synthesize the information contained in the T-S diagrams with the possibility of reducing the size of the representation, since only few modes are able to capture the vertical variability in the ocean. This means that, for open ocean conditions and long time scales (from synoptic to seasonal), it is possible to reduce the vertical complexity or degrees of freedom of a dynamical state variable. In other words, if $\Phi(x, y, t, z)$ is the dynamical variable, and we can separate it as $\Phi(x, y, t, z) = \Sigma_i \alpha_i\ (x, y, t)\ e_i(z)$, then only a reduced number of e_i could be used if the basis functions are vertical EOFs.

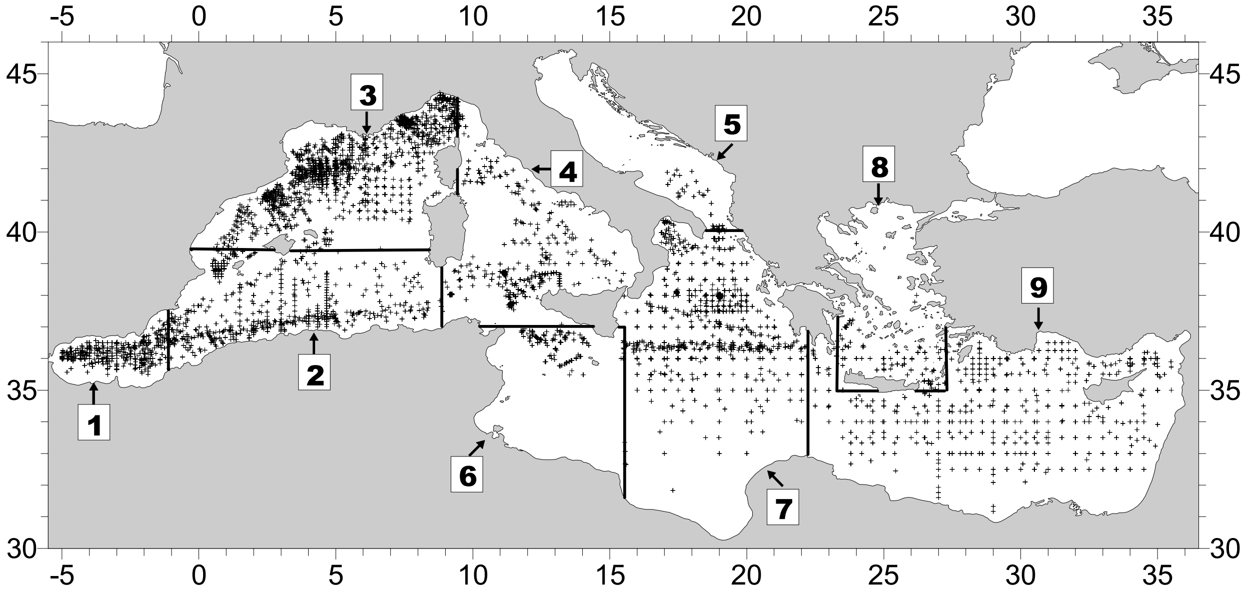


Fig. 1. Regional breakdown of the Mediterranean Sea with the positions of Medatlas casts indicated by crosses in the plot. The regions are the following: **(1)** Alboran Sea; **(2)** Algerian Basin; **(3)** northwestern Basin; **(4)** Tyrrhenian Sea; **(5)** Adriatic Sea; **(6)** Strait of Sicily; **(7)** Ionian Sea; **(8)** Aegean Sea; **(9)** Levantine Basin.

This important discovery has led recently to the development of reduced order data assimilation techniques, such as the Scheme for Ocean Forecasts and Analysis (SOFA, De Mey, 1997) used in MFSPP. Such a scheme projects the difference between observations and model first guess (misfit) into vertical EOFs and reduces the OI technique to a two-dimensional instead of a three-dimensional multivariate problem. However, EOFs should be checked carefully and should be significant if they are to play such an important role in the assimilation.

For the Mediterranean Sea, vertical EOF studies have been undertaken for isolated regions with in situ data (Hecht et al., 1988; Nittis et al., 1993) and for model simulations (Korres et al., 2000b). In these studies, it was again shown that a small number of vertical modes could explain most of the vertical variance in the temperature and salinity profiles, and even fewer modes could represent the dynamic height vertical structure. The vertical EOFs were also interpreted in terms of known water masses but the in situ data were very limited on the time and space scales. A consistent analysis of an extensive Mediterranean in situ data set has not been done before this work.

In this paper, optimal sets of bivariate EOFs for the combined temperature and salinity variability in the Mediterranean basin are calculated and shown to be able to efficiently reproduce the statistics of the vertical thermohaline structure of the basin. Particular attention has been paid to study the "significant" bivariate EOFs and "optimal" reduced set of bivariate EOFs. Only the water column from the surface to 480 m was investigated, which is the standard sampling depth limit for most XBTs.

The paper is organized as follows. In Sect. 2, we present the in situ and model data sets, and the method used to calculate climatologies. In Sect. 3, we describe the methodology used for the EOF calculations and the methods adopted for defining the "optimal" reduced EOF subspace and the "significant" EOFs. In Sect. 4, we discuss the seasonal variability of the in situ and model simulation data sets. In Sect. 5, we present the bivariate EOF structure from both in situ and model simulation data. In Sect. 6, we compute the optimal set of reduced modes capable of reproducing the observed temperature and salinity profiles. Finally, a summary of the results and the conclusions are presented in Sect. 7.

2 The data bases

2.1 In situ data

The in situ data used in this study are temperature and salinity profiles from the Medatlas hydrological database prepared by a consortium of several Mediterranean data centers following a common protocol (Fichaut et al., 1998). The protocol includes a quality control procedure based on the recommendations of the Intergovernmental Oceanographic Commission (IOC) and the European Marine Science and Technology Program (MAST). Data quality checks consisted of controlling whether individual values fell within the specific minimum and maximum values defined for recognized rectangular sub-regions (see Fig. 1 in Fichaut et al., 1998), and evaluating the behaviour of the same values with respect to accepted limits based on pre-existing statistics obtained from the LEVITUS climatology (Levitus et al., 1994) and the MODB climatology (Brasseur et al., 1996).

CTD and bottle data acquired from 1970 to 1995 were extracted from Medatlas; only data passing the imposed quality

Table 1. Temporal breakdown by season, by month, and by region of the in situ profiles from the Medatlas data set whose minimum depth is 480 m

REGION	1	2	3	4	5	6	7	8	9
Jan	9	20	45	7	0	2	17	1	1
Feb	9	22	324	11	1	3	16	2	25
Mar	31	32	329	117	8	47	28	36	34
Winter	49	74	698	135	9	52	61	39	60
Apr	70	31	206	17	6	0	72	12	99
May	71	33	131	25	9	0	18	10	31
June	25	203	187	16	0	15	52	0	2
Spring	166	267	524	58	15	15	142	22	132
July	23	84	223	79	4	11	70	4	16
Aug	3	14	36	29	4	12	15	3	31
Sept	39	38	44	35	1	14	136	13	44
Summer	65	136	303	143	9	37	221	20	91
Oct	72	58	64	62	6	24	70	8	149
Nov	21	41	78	12	1	22	236	5	80
Dec	7	22	83	28	4	0	14	3	0
Autumn	100	121	225	102	11	46	320	16	229
Grandtotal	380	598	1750	438	44	150	744	97	512

Table 2. Temporal breakdown by season and by region of the simulated profiles whose minimum depth is 480 m

REGION	1	2	3	4	5	6	7	8	9
Winter	764	4195	3485	3391	441	858	7388	1069	9036
Spring	774	4210	3446	3416	423	886	7534	1073	8995
Summer	768	4161	3465	3419	434	872	7568	1057	8849
Autumn	760	4180	3465	3405	443	871	7474	1045	8966

checks (indicated by a control flag equaling 1) were selected providing a total of 22 268 profiles. Our analysis was done in order to study the upper thermocline variability, so the maximum depth selected for our profiles is 480 m which contains most of the seasonal, mesoscale and interannual variability signal. The original profiles were linearly interpolated at 16 standard levels (5, 15, 30, 50, 70, 90, 120, 160, 200, 240, 280, 320, 360, 400, 440, 480 m) and reorganized into 9 smaller data sets, one for each region shown in Fig. 1. The first and foremost criterion we adopted to divide the Mediterranean Sea into sub-regions had a physical basis, that is the known variability associated with the dynamical regimes peculiar to different areas, following Pinardi and Masetti (2000). The second criterion is purely numerical and its application ensures that enough data are available for the EOF calculation in each region. Essentially, the 9 regions we identified thus satisfy these two requirements. We are aware that a more detailed subdivision of the Mediterranean Sea would have been better from a purely oceanographical point of view, and that this would have produced different EOFs, but the actual distribution of in situ data unfortunately does not permit this.

Each regional data set was then carefully checked, in order to eliminate all the profiles containing density inversions in the vertical, to ensure vertical stability. Then, unreasonable data were identified level by level as those exceeding the seasonal mean by more than 3-standard deviations and eliminated. Finally, incomplete profiles, namely those containing gaps in the vertical or shallower than 480 m of depth were discarded. At the end of this processing, the set of observations shrunk to a total of 4713 profiles. The distributions by season, by month, and by region of the final set of observations are indicated in Table 1 and the crosses in Fig. 1 show their geographical location. The uneven distribution in both time and space is clearly notable.

2.2 Model data

The model data used are monthly mean fields of temperature and salinity from a model simulation done for the period January 1979 – December 1993. The model simulation is documented in Demirov and Pinardi (2002), and here we will describe only a few of its characteristics.

The model used in MFSPP is the Modular Ocean Model (MOM), which was adapted to the Mediterranean Sea by Roussenov et al. (1995). The model grid has 31 vertical levels, as in Korres et al. (2000a), and a horizontal resolution of $1/8° \times 1/8°$, which does not include the northern part of the Adriatic Sea. Horizontal turbulent mixing is biharmonic, with tracer coefficients equal to 1.5×10^{10} m^4 s^{-1} and momentum coefficients equal to $2 \times 10^{10} \sim$m^4 s^{-1}. Vertical turbulent processes are parameterized by a constant turbulent diffusion coefficient set to 0.3×10^{-4} m^2 s^{-1} and a viscosity coefficient equal to 1.5×10^{-4} m^2 s^{-1}. A standard convective adjustment procedure (Cox, 1984) is applied when static instability appears in the water column. This vertical mixing scheme choice has been studied in the previous work of Korres et al. (2000a) and Castellari et al. (2000).

The transport through the Strait of Gibraltar is parameterized by extending the model area westward of Gibraltar up to a longitude 9.25°W. In this Atlantic box, lying between latitudes $33°30'$ N and 37° N, the surface forcing is switched off, and the temperature and salinity are relaxed towards annual mean climatological fields.

The surface forcing is computed in an interactive way with 6-hourly ECMWF (European Center for Medium Range Weather Forecast) atmospheric reanalysis fields and Sea Surface Temperature (SST) from the model. The different components of the surface net heat flux are computed on the basis of the parameterizations of Reed (1977) for the surface solar radiation flux, of Bignami et al. (1995) for outgoing, longwave radiation, and of Kondo (1975) for sensible and latent heat fluxes. The bulk formulation of Hellerman and Rosenstein (1983) is used in the wind stress computation. Descriptions of the implementation and test of the surface momentum and heat flux parameterizations can be found in Castellari et al. (1998, 2000). The meteorological data used are the atmospheric temperature and humidity at 2 m and wind components at 10 m. Cloud cover is taken from the monthly mean COADS data (Da Silva et al., 1994). The sea surface water flux is parameterized with a salt flux given by the relaxation of model sea surface salinity towards a new climatology called MED6 (Brankart and Pinardi, 2001), that we will better describe in the following section. The relaxation constant is 2 m/days everywhere.

To reduce the amount of data to process, the original data set was subsampled at 1/2°×1/2° horizontal resolution. The remaining data set was pre-processed with the same procedure used for the in situ data set, in order to eliminate the unreasonable and critical values. The distribution by season and by region of the data suited to the study is shown in Table 2. Obviously, in this case, the data distribution is regular both in space and in time.

2.3 Climatology

The temperature and salinity fields that will be used in our calculations are the departures of the observed or simulated variables from some seasonal climatology. We have two possible choices for the climatology: a regional mean profile, that is the average of all the profiles available in a given region, or a gridded climatology, computed by objective mapping. The first is the standard approach in this type of calculation and it is widely used in literature (e.g. Fukumori and Wunsch, 1991; Gavart and De Mey, 1997; Maes, 1999). The second was used first in Faucher et al. (2002) and it looks more promising since it considers the spatial variability of the mean within each region. Whatever the choice, when calculating the departure from monthly mean climatologies, we will obtain anomalies containing information at both the interannual and the mesoscale time frequencies. Due to scarcity of data, we cannot distinguish between the two frequencies and we will consider the full anomaly signal. Thus, we decided to subtract from the original data the monthly gridded climatology.

Different climatologies were used for in situ and model data calculated from each of the two data sets. The model data climatology was simply calculated by averaging all the profiles resulting at different times at each grid point, and grouping them by month. Regarding the in situ data, we used the monthly averaged MED6 climatology with a horizontal resolution of 0.25° calculated from the Medatlas data set (Brankart and Pinardi, 2001), applying an improved objective analysis technique already used for calculating the MED2 climatology (Brasseur et al., 1996).

3 Methods

3.1 Multivariate vertical EOF

The EOF analysis and its equivalent formulation, the Principal Components analysis, are tools widely used in atmospheric science and oceanography (see, for instance, Lorentz, 1956; Preisendorfer, 1988; Fukumori and Wunsh, 1991; and many others). One of its common applications is in reducing the dimensionality of a problem, and in transforming interdependent coordinates into significant and independent ones (e.g. De Mey, 1997; De Mey and Benkiran, 2002). Our calculation is based on a bivariate approach that isolates the primary modes of the combined variance of temperature and salinity profiles (e.g. Gavart and De Mey, 1997; Maes, 1999).

Without entering into the details of the method, we will hereupon give a summary of the procedure we adopted so as to clarify the notations and conventions that were employed specifically by us. Following Gavart and De Mey (1997), any hydrological temperature and salinity profile is transformed into a state vector $\boldsymbol{x}$ containing the $2M$ variables needed to describe the ocean thermohaline state in the vertical, that is:

$$\boldsymbol{x} = [x_1, \ldots, x_{2M}] = \left[\frac{\delta T_1}{\sigma_1^T}, \ldots, \frac{\delta T_M}{\sigma_M^T}, \frac{\delta S_1}{\sigma_1^s}, \ldots, \frac{\delta S_M}{\sigma_M^S} \right], (1)$$

where M is the number of vertical levels and $\delta T_k = (T_k - T_k^{\text{clim}})$ and $\delta S_k = (S_k - S_k^{\text{clim}})$, $k = 1, \ldots M$, are the departures or anomalies from the climatology at each vertical level, normalized by their standard deviation from climatology, σ_k:

$$\sigma_k^T = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(T_k - T_k^{\text{clim}} \right)_n^2} \,,$$

$$\sigma_k^s = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(S_k - S_k^{\text{clim}} \right)_n^2} \quad k = 1, \ldots M \,. \tag{2}$$

Here, N is the number of observations at each level (i.e. the number of vertical profiles). Different normalization factors (2) were calculated for each vertical level. In doing so, we ensure that the first EOF modes will capture the bulk of the variability in the section of the water column that was considered, thereby representing the variances relating to its upper and deeper portions as well. Alternatively, we could have used uniform (in the vertical) normalization factors, but in this case, the first EOF modes would have represented mostly the variability associated with the upper levels which are characterized by a higher variance.

We obtained the vertical EOFs as the eigenvectors of the covariance matrix, **C**, calculated over the set of realizations of $\boldsymbol{x}$ (the N profiles) using the MATLAB® function pcacov. The amount of variance accounted for by each eigenvector, $\boldsymbol{e}_i$, was calculated, as usual, as the Percentage of Variance Explained (PVE):

$$PVE_i = \frac{100 \cdot \lambda_i}{\sum_{k=1}^{2M} \lambda_k} . \tag{3}$$

Since the first eigenvalue has normally the largest value, the first mode accounts for the largest variance in the water column.

The complete set of eigenvectors above forms an orthonormal and complete basis in which we can represent each state vector according to

$$\boldsymbol{x} = \sum_{i=1}^{2M} a_i \boldsymbol{e}_i \,, \tag{4}$$

where the a_i are the amplitudes or scores or principal components.

3.2 Reducing the order of the EOF space: the optimal set of EOF

Generally, a limited number of EOFs are able to describe most of the variance in a data set, in particular those with the largest eigenvalues. This reduced set of EOFs can be used to represent efficiently the physical field under study, and we call this subset the "optimal set". The capacity to represent a physical state in a reduced order form, preserving its physical significance, is a fundamental requirement in practical assimilation problems, where the order reduction could diminish significantly the size of the algebraic calculations involved (De Mey, 1997; De Mey and Benkiran, 2002).

There are, however, several methods to decide what is the "reduced and/or optimal" set of EOFs, and we will list here the few that we have used. Provided we have identified the $m < 2M$ dominant modes, the state vector can be represented by the truncated form,

$$\boldsymbol{x}_r \cong \sum_{i=1}^{m} a_i \boldsymbol{e}_i \tag{5}$$

Among all the possible basis functions, the EOFs are the optimum set for a given truncation m because they minimize the residual variance (Lorenz, 1956), i.e.:

$$\boldsymbol{R} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{r}_n^2 = \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{x} - \boldsymbol{x}_r)_n^2 \;. \tag{6}$$

As before, N is the number of observations at each depth, the latter being indicated by the vector of dimension 2M. The square root of $\boldsymbol{R}$ is also referred to as the root mean square residual error (e.g. Fukumori and Wunsch, 1991).

The most common and simplest approach to establish the optimum truncation index is described in Preisendorfer (1988; page 192). One first examines the sequence of eigenvalues, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{2M}$, and looks at their magnitudes. Occasionally, after a certain index, the values drop abruptly and remain relatively small. The sharp change of slope in the eigenvalue curve indicates that the last eigenvectors are trying to fit the noise in the given profiles. The index value corresponding to a large drop in λ values can hence be chosen as the truncation index m in Eq. (5). Concurrently, the structure of the first m eigenvectors looks simpler and less "noisy" than the structure of the higher modes. Preisendorfer (1988) also listed a number of selection rules based on statistical hypotheses to separate signal and noise eigenvalues. Of these, "Rule N", a dominant-variance rule that is based on the premise that the larger-variance terms are associated with the signal, is often used (e.g. Preisendorfer et al., 1981; Frankignoul and Reynolds, 1983; Korres et al., 2000b). This rule simulates random sampling of data from a normal population – the noise subspace – by means of a Monte Carlo procedure, and builds up a cumulative distribution for each eigenvalue. By testing the null hypothesis that the data sample has been drawn from the normal population, it identifies the significant eigenvalues at the 5% level of significance as the ones larger than the 95% point on the cumulative distribution of the noise spectra. Due to its statistical nature, this selection rule could fail when the sample size is small (as it is with our in situ data set), when the data are highly non-normally distributed, or when correlations in the data set are large. Thus, we adopted a further criterion based on the comparison between the vertically averaged standard deviations from climatology obtained as a weighted vertical average of (2):

$$\overline{\sigma^T} = \sum_{k=1}^{M}\left(p_k \sigma_k^T\right) \Big/ \sum_{k=1}^{M} p_k \quad \overline{\sigma^S} = \sum_{k=1}^{M}\left(p_k \sigma_k^S\right) \Big/ \sum_{k=1}^{M} p_k \tag{7a}$$

and the weighted vertical average of the root mean square residual, defined using Eq. (6),

$$\begin{aligned} \overline{rms(\delta T)} &= \sum_{k=1}^{M}\left(p_k \sigma_k^T \sqrt{R_k}\right) \Big/ \sum_{k=1}^{M} p_k \\ \overline{rms(\delta S)} &= \sum_{k=1}^{M}\left(p_k \sigma_k^S \sqrt{R_k}\right) \Big/ \sum_{k=1}^{M} p_k \end{aligned} \tag{7b}$$

where the weights p_k-s are calculated as the difference between adjacent depths and the sum is extended over the full set of vertical levels. The criterion consists of choosing the index m such that the vertical averaged residual rms is much less than the observed vertically averaged standard deviation. This means that the ratio of Eq. (7b) to Eq. (7a) should be small and in fact, the truncation index was chosen in such a way that this ratio had to be less than 0.3, a threshold value which ensures that a variance greater than 90% is explained. Considering that the levels close to the surface have the highest variance, the criterion is more rigorous if it is applied by partitioning the water column into two parts: from 0 to 200 m and from 240 m to 480 m. Thus, the value of $rms(\delta T)$ and $rms(\delta S)$ were calculated for the above two layers separately, as a function of an increasing truncation index.

3.3 Statistical significance of EOFs

Most of the papers involving EOF calculations in atmospheric and oceanic science associate the concept of statistical significance with the operation of separating the subset of "significant" eigenvalues that constitute the signal-space from that which constitutes the noise-subspace, with the objective – discussed in the previous paragraph – of reducing the dimensionality of a problem (e.g. Korres et. al., 2000b) or filtering the data (e.g. Frankignoul and Reynolds, 1983). The procedures adopted in these cases are statistical tests, often based on Monte Carlo simulations, that generate the probability distribution necessary to identify the group of "significant" eigenvalues by testing a null hypothesis at some prescribed level of statistical significance. The already mentioned Rule N (see Sect. 3.2) is an example. Apart from their utility in identifying a truncation point for the reduced set of EOFs, these procedures do not really address the question of the quality of the estimation of the EOF patterns. Moreover, it is difficult, if not impossible, to build the required probability distribution when the available sample size is small, as it is with our in situ data set.

The evaluation of the quality of the estimated EOF is an important issue, especially when the sample size is small and, therefore, provides only a rough estimate of the covariance matrix. It obviously follows that the calculated EOFs will also be only a rough estimate of the true ones. Von Storch and Hannoschöck (1986) showed that the sample eigenvalues are biased estimators of the true eigenvalues, the bias being inversely proportional to the number of independent samples. Generally, the largest eigenvalues are overestimated and the smallest ones are underestimated. Consequently, the sample variance expressed by the corresponding EOF is systematically an over- or under-estimation of the true variance.

Moreover, sampling errors occur when the available sample is too small. North et al. (1982) demonstrated that when the eigenvalues are closely spaced, the corresponding EOFs could be a bad approximation of the true ones, if the sampling error is comparable to the spacing of the eigenvalues. They formulated a rule-of-thumb to evaluate the sampling errors involved in the EOF calculations and thereafter, to determine if a particular EOF is significantly different from its neighbour. This rule is very easy to apply and suitable for our study. So we decided to adopt it as an empirical test of statistical significance for our calculations. Some more details on this rule are given below.

Using standard linear analysis arguments, North et al. (1982) evaluated the "typical errors" of the eigenvalues and eigenvectors as:

$$\delta\lambda_k \approx \lambda_k \cdot \sqrt{\frac{2}{N}} \tag{8a}$$

$$\delta \boldsymbol{e}_k \approx \frac{\delta\lambda_k}{\lambda_j - \lambda_k} \cdot \boldsymbol{e}_j, \tag{8b}$$

where λ_j is the eigenvalue closest to λ_k and N is the number of independent samples. The following "rule-of-thumb" was then formulated:

"If the sampling error, $\delta\lambda_k$, of a particular eigenvalue is comparable to or larger than the spacing between λ_k and a neighboring eigenvalue, λ_i, then the sampling error $\delta\boldsymbol{e}_k$ for the EOF associated with λ_k will be comparable to the size of the neighboring EOF, $\boldsymbol{e}_j$".

Consequently, a kind of degeneracy occurs for which the two EOFs are intrinsically ambiguous; in fact, any linear combination of them is also an eigenvector. When this happens, different sampling methods will lead to drastically different EOFs, depending on the combinations between degenerate EOFs that are picked up. It follows that the particular EOF is not properly resolved and caution is called for when trying to provide a physical interpretation. The occurrence or not of degeneracy depends on the quantity of data processed: the more data, the smaller the "typical errors" in Eqs.(8a, b).

4 Analysis of the data sets

4.1 The climatology

In this section, we will discuss the characteristics of the climatology used for the computations of anomalies. The seasonal characteristics of the climatological temperature and salinity fields are displayed in Figs. 2 and 3 for the in situ and the model data. We show the temperature and salinity fields corresponding to the depths of 50 and 280 m for winter and summer, where winter is the average of January, February and March, and summer is the average of June, July and August. The two depths above were chosen as representative of the surface and intermediate layers, where two different Mediterranean water masses are found. The surface water layer contains the so-called Modified Atlantic Water (MAW) due to its origin and successive modifications induced by mixing processes. MAW occupies typically the first 100–200 m, and is characterized by low salinity values. How low these values are depends on the geographical location: the closer the sampling location is to Gibraltar, the fresher and more superficial this layer is. The second layer is representative of the Levantine Intermediate Water (LIW), which is formed in the eastern Mediterranean during winter and spreads throughout the Mediterranean at intermediate depths. The LIW is characterized by a subsurface salinity maximum that is higher and shallower near the formation region. Since the cores of these water masses occupy different depths in several parts of the basin, the chosen depths are only partially representative of them.

Comparing Figs. 2 and 3, it is evident that the climatology simulated by the model and MED6 are rather alike. Similar features are observed on the basin scale, but differences exist in regions where data scarcity is larger. This is the case for the salinity field at 280 m in the southern Ionian Sea, where a patch of low salinity waters is present, probably due to the extrapolation done by the objective analysis technique. On the other hand, the model shows a drift toward a fresher and warmer climate due to unresolved physical processes. Model drift has been documented in several papers during the last few years (Roussenov et al., 1995; Demirov and Pinardi 2002). Generally, it is due to surface forcing inaccuracies that produce water masses which are different from those that are present in reality, and to the Gibraltar Strait parameterization.

Both the temperature and salinity distributions show a positive west-east gradient. In winter, the temperature difference between west and east is 4°C at 50 m depth in both climatologies (Figs. 2a and e), but the absolute values are 0.2°C higher in the model than in MED6. In summer, the same gradient is 7°C and the difference between model and in situ values

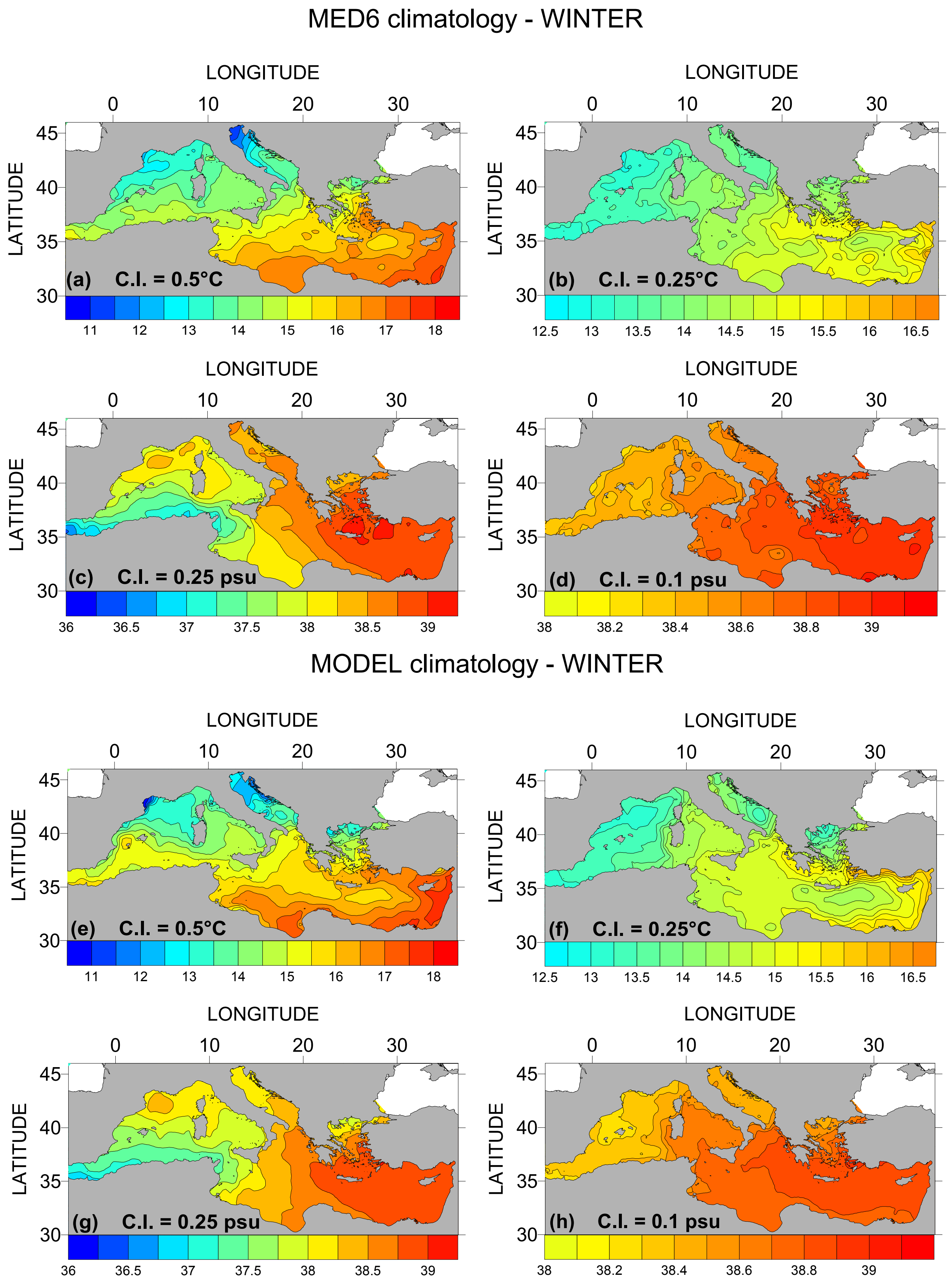


Fig. 2. Winter MED6 and model climatological fields at 50 m and 280 m.

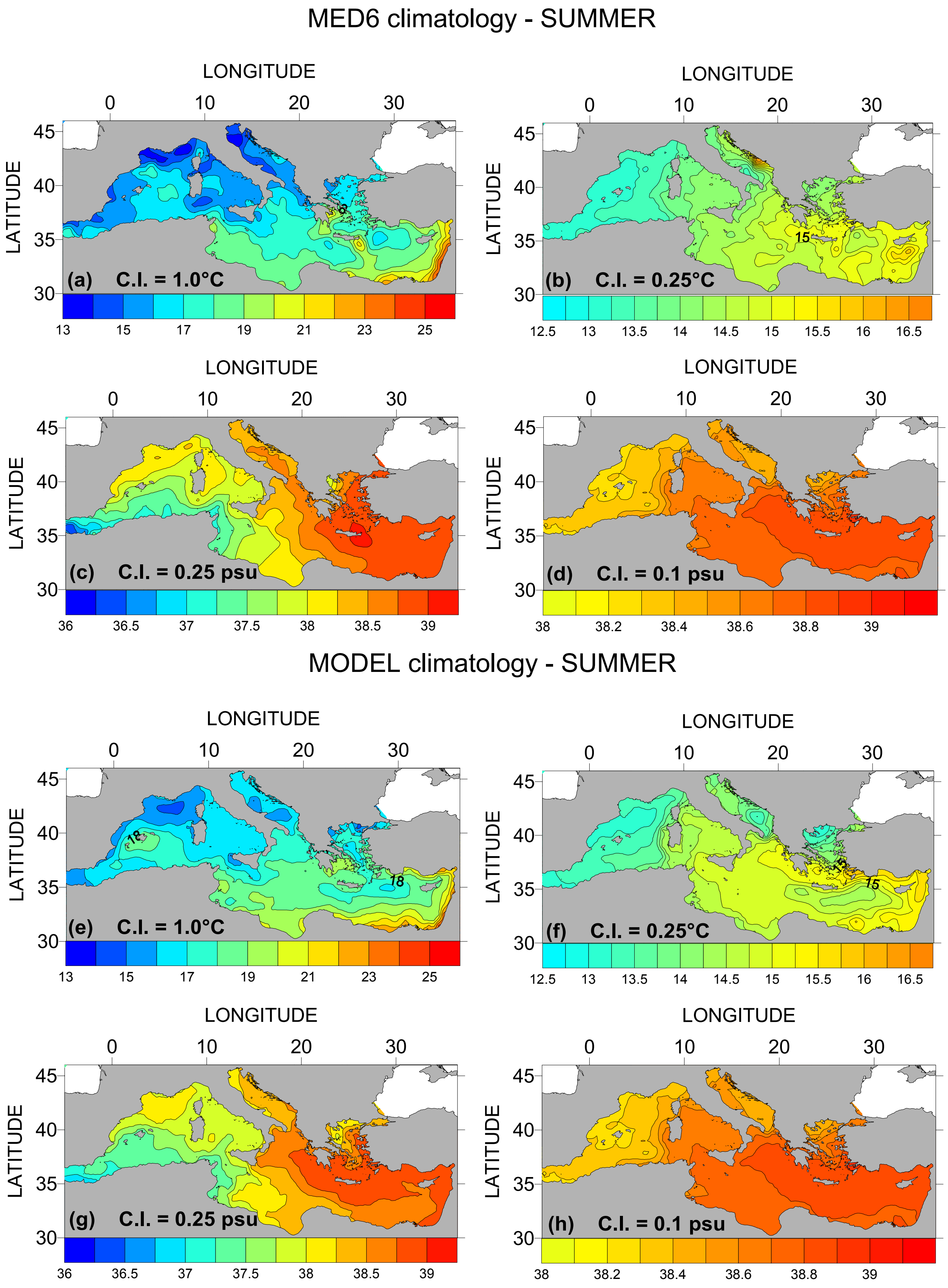


Fig. 3. Summer MED6 and model climatological fields at 50 m and 280 m.

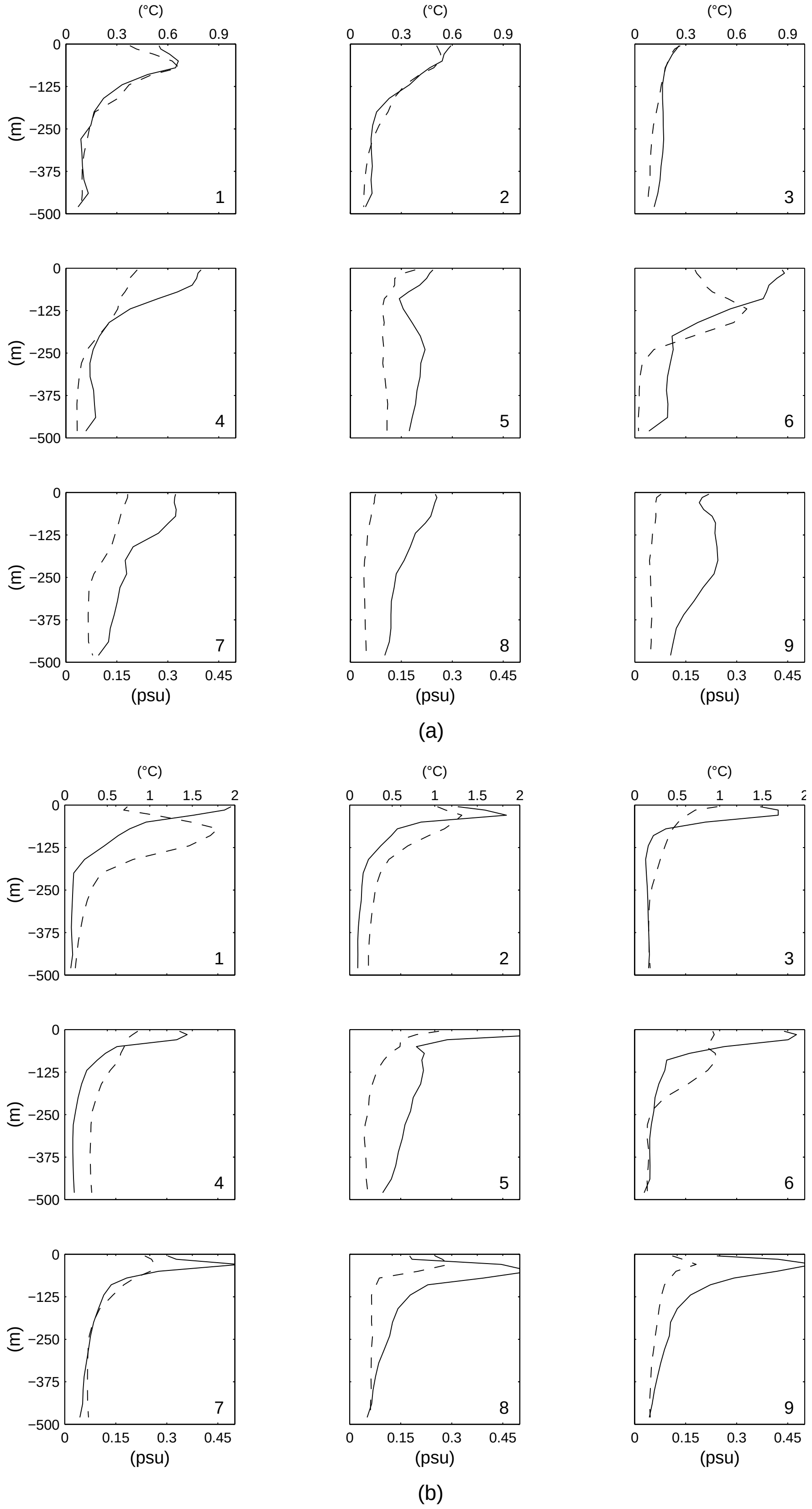
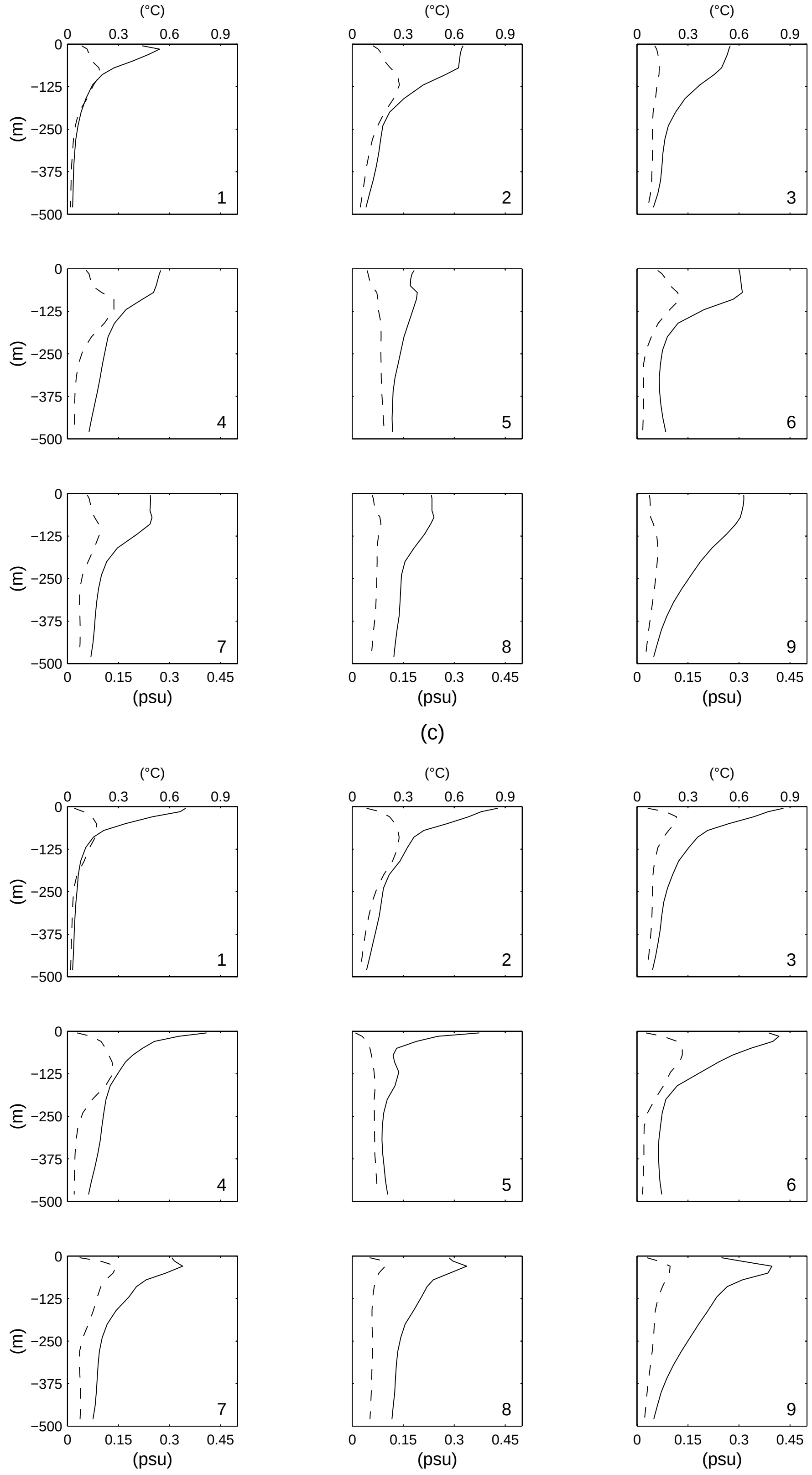


Fig. 4. Vertical profiles of the standard deviations from climatology for temperature (solid curve) and salinity (dotted curve). The in situ data for the nine regions (indicated in the lower right corner) are shown for winter **(a)** and summer **(b)**. Model simulation results are shown in **(c)** for winter and **(d)** for summer (Fig. 4 continues).



(c)

(d)

Fig. 4. ... continued.

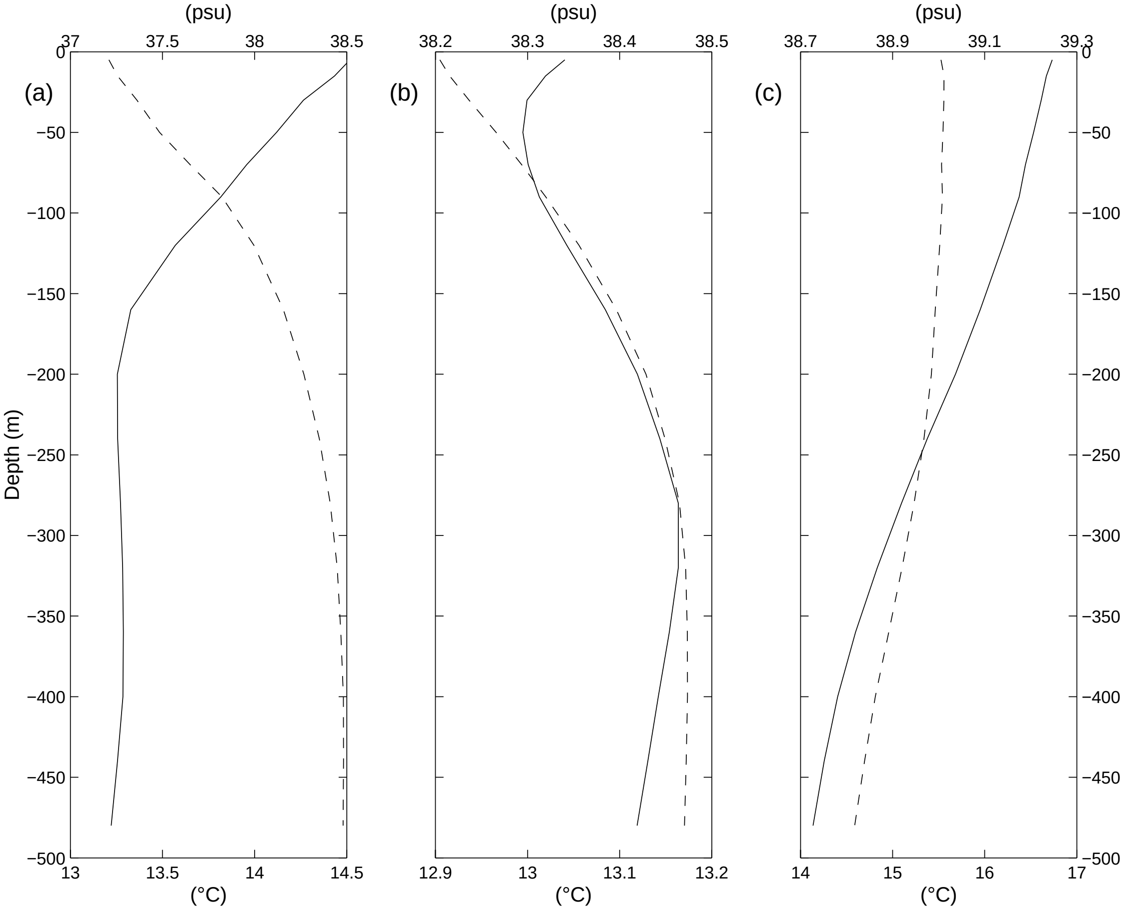


Fig. 5. Mean winter vertical structure of temperature (solid line) and salinity (dashed line) calculated as regional averages for **(a)** Algerian Basin, **(b)** northwestern Basin and **(c)** Levantine Basin.

increases up to 1°C (Figs. 3a and e). The seasonal variations, affecting predominantly the surface, are greater in the eastern than in the western Mediterranean, partially due to the lower latitudes of the eastern basin and the different air-sea interaction processes that are occurring. At the intermediate level, the Ionian and Tyrrhenian Seas are found to be warmer in the case of the model data than in the case of the in situ data by, respectively, 0.3–0.6°C and 0.3°C in winter (Figs. 2b and f) and 0.3°C and 0.6°C in summer (Figs. 3b and f).

The salinity gradient between west and east is not affected as much by the seasonal changes as the temperature gradient is. Comparing the salinity fields at 50 m, we note a general agreement between the results obtained with the model and the in situ data both in winter (Figs. 2c and e) and in summer (Figs. 3c and e). At the depth of 280 m, the model data is highly smoothed when compared to the in situ data, and shows the salinity to be slightly lower in the eastern part (compare Figs. 2h and d and Figs. 3h and d). Even if differences do exist, we argue that the two climatologies are close enough to allow the comparison of the EOFs resulting from the anomalies.

4.2 Vertical variability of temperature and salinity

Vertical profiles of the standard deviation from climatology, described by Eq. (2), are shown in Fig. 4 for both in situ and model data. Only the results for summer and winter are shown, representing the two extremes of the stratification regime, the first dominated by a well stratified surface layer and the second by vertical homogenization. The results for the Adriatic Sea when using in situ data are notably uncertain due to the data scarcity (see Table 1).

Looking first at the calculation from in situ data (Figs. 4a and b), we observe that the largest variability is found in the surface layers in most regions. Typical values of the temperature and salinity departures from climatology are, respectively, around 0.6°C and 0.2 psu in winter and 2°C and 0.3 psu in summer. During winter (Fig. 4a), due to the intense vertical mixing processes occurring in the basin, most of the regions show a surface maximum of T variability between 0 and 100 m. However, four regions show exceptions to this rule. The first, region 1, shows the subsurface T and S variability maxima, due to the different degrees of mixing that the Atlantic water is subjected to entering the Gibraltar Strait. The second is the Strait of Sicily (region 6), where we observe a subsurface maximum of salinity variability probably due to the amount of LIW that is able to spread across the Strait. The third exception is the T variance signal in the Levantine basin (region 9), where intermediate water formation occurs, giving rise to a 250 m subsurface maximum in variance. Finally, small variations are found in the northwestern Basin (region 3) with similar magnitudes in the case of both the surface and the intermediate layer. Temperature variations are about 0.3°C at the surface and 0.1°C at 500 m, while the salinity variability ranges from 0.1 to 0.05 psu. This is due to the vertical homogenization processes occurring in this region, which result in a low standard deviation in property values for the kinds of water masses formed. During summer (Fig. 4b), the variability is dominated by the seasonal thermocline depth variations that produce a subsurface variability maximum around 30–50 m, i.e. the depth of the summer surface layer. The variations in the intermediate layer are about 0.1°C and 0.04 psu both in winter and in summer.

Model data (Figs. 4c and d) give similar information regarding the vertical structure, but the order of the variations are generally smaller than those obtained from the direct observations. This is particularly evident at the surface, where the maximum variation in summer is only 0.7–0.9°C for temperature and 0.15 psu for salinity. On the contrary, region 3 shows higher variability during the winter (Figs. 4c and a), probably due to the inability of the model to form enough deep waters in this season.

The comparison between the in situ data and the model data calculations suggests that the model is a sort of filtered representation of the variability contained in the in situ data.

5 The seasonal bivariate vertical EOFs

5.1 Results from in situ data

In this section, we will discuss the EOF analysis of in situ data focussing on three regions, namely the Algerian Basin, the northwestern Basin and the Levantine Basin. These are regions where the circulation and the associated mesoscale phenomena are particularly important for the whole Mediterranean. In Fig. 5, we show the three regional mean winter vertical profiles of temperature and salinity. In the Algerian Basin, the surface water is characterized by a minimum averaged salinity of 37.2 (Fig. 5a), which in the northwestern Basin is 38.2 (Fig. 5b). The underlying layers are occupied by LIW, characterized by a salinity of about 38.48 in the Algerian Basin and one ranging between 38.4–38.5 in the northwestern Basin. In the Levantine basin, LIW occupies a surface layer and its averaged salinity is about 38.9 (Fig. 5c).

As explained in Sect. 3.1, all the calculations are done using dimensionless variables. To aid in understanding and interpretation, the results presented here use the dimensional form of the variables. Dimensional forms are obtained by multiplying the variables by the corresponding normalization factor. The same is done for the modes that are furthermore multiplied by the respective eigenvalue so that they can be compared to each other and interpreted in terms of their contribution to the total variance.

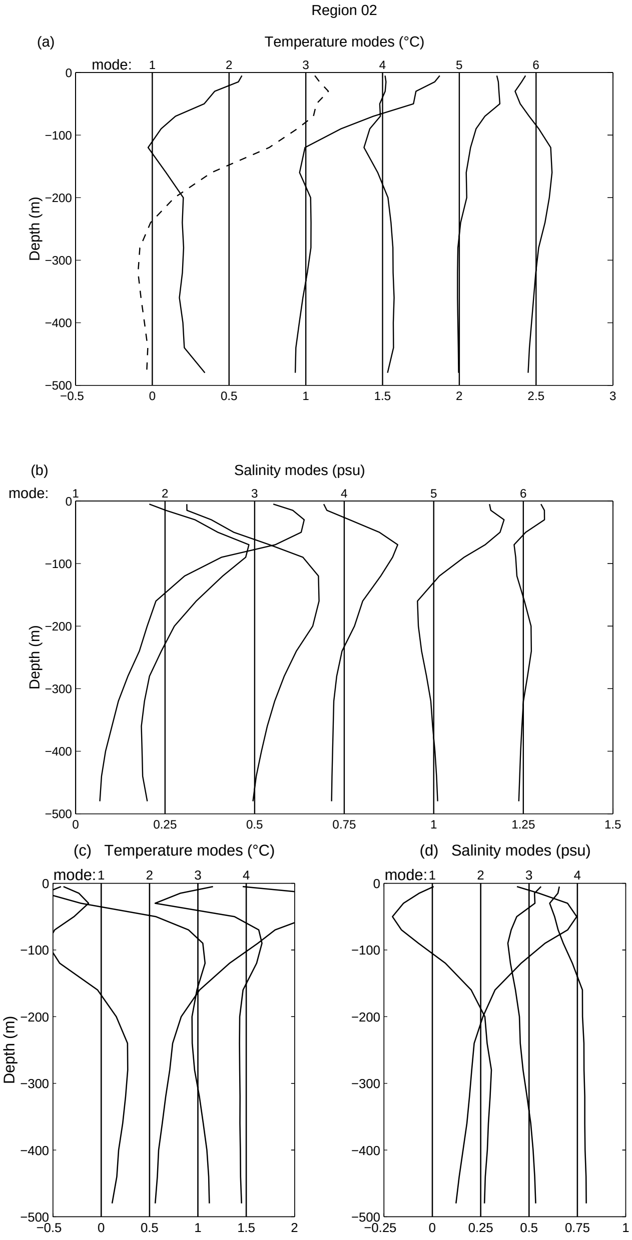


Fig. 6. Vertical EOFs for region 2 (Algerian Basin): **(a)** temperature and **(b)** salinity modes in winter, and **(c)** temperature and **(d)** salinity modes in summer. The first six winter modes and four summer modes are shown. For graphical convenience, a dashed line is used to indicate a change in sign, and respective offsets of 0.5°C and 0.25 psu have been added to each T and S mode for ease of presentation.

5.1.1 Algerian Basin

Figures 6a and b show the first six vertical modes for both temperature and salinity calculated from the winter data set. For graphical convenience, the sign of the first mode in Fig. 6a was changed, and a dashed line has been used to

Table 3. The first ten eigenvalues of the covariance matrix of the in situ data in region 2 for winter and summer. The first column contains the eigenvalues. The percentage of the variance explained (PVE, Eq. (3) is shown in the second column. In the remaining two columns for each season, the ratio of the vertical average of the root mean square residual (7b) to the vertically averaged standard deviation (7a) is also shown. This latter is shown only for the layer between 0 and 200 m

	Winter				Summer			
Mode	λ	PVE	rms ΔT (°C)	rms ΔS	λ	PVE	rms ΔT (°C)	rms ΔS
		(%)	0.356 °C	0.183		(%)	0.570°C	0.194
1	10.3	31.6	0.79	0.75	11.6	36.0	0.98	0.68
2	7.9	24.4	0.67	0.71	7.6	23.5	0.70	0.53
3	4.7	14.6	0.53	0.57	3.2	10.0	0.63	0.47
4	2.6	8.0	0.51	0.50	2.7	8.3	0.59	0.43
5	2.1	6.3	0.45	0.35	1.9	6.0	0.50	0.38
6	1.4	4.3	0.38	0.33	1.2	3.9	0.44	0.36
7	0.8	2.4	0.34	0.29	0.8	2.6	0.37	0.34
8	0.7	2.2	0.27	0.27	0.72	2.3	0.31	0.30
9	0.44	1.4	0.22	0.25	0.45	1.4	0.27	0.26
10	0.36	1.1	0.20	0.20	0.37	1.2	0.27	0.20

Table 4. The first ten eigenvalues of the covariance matrix of the in situ data in region 3 for winter and summer. The first column contains the eigenvalues. The percentage of the variance explained (PVE, Eq. (3) is shown in the second column. In the remaining two columns for each season, the ratio of the vertical average of the root mean square residual (7b) to the vertically averaged standard deviation (7a) is also shown. This latter is shown only for the layer between 0 and 200 m

	Winter				Summer			
Mode	λ	PVE	rms ΔT (°C)	rms ΔS	λ	PVE	rms ΔT (°C)	rms ΔS
		(%)	0.181°C	0.084		(%)	0.469°C	0.102
1	12.8	39.9	0.85	0.90	13.7	42.5	0.90	0.71
2	6.8	21.2	0.74	0.52	6.1	19.1	0.86	0.52
3	4.7	14.5	0.51	0.50	3.2	10.0	0.77	0.47
4	2.5	7.9	0.42	0.40	2.6	7.9	0.55	0.46
5	1.50	4.7	0.40	0.32	1.3	4.1	0.53	0.40
6	0.88	2.7	0.33	0.30	1.00	3.1	0.50	0.37
7	0.68	2.1	0.29	0.27	0.92	2.9	0.41	0.35
8	0.47	1.5	0.26	0.24	0.69	2.1	0.35	0.31
9	0.36	1.1	0.24	0.21	0.53	1.7	0.31	0.28
10	0.26	0.8	0.22	0.18	0.37	1.2	0.26	0.25

indicate the inversion. A first look shows that most of the signal is confined above 250 m for all the eigenvectors. In Table 3, we show the PVE for each mode and two seasons. The contribution to the variance of the upper portion of the water column is given mainly by the first three modes that together account for 71% of the total variance. Mode 1 (31%) displays maximum fluctuations within the first 70 m from the surface. Therefore, this mode probably represents mostly the variability of the mixed layer. A further contribution to the variability of this layer comes also from mode 3 (15%). Mode 2 (24%) shows two separate maxima for T and S, respectively. A maximum amplitude for T anomalies is found around 120 m, in correspondence to a maximum in the S anomaly displayed by the third mode, while the maximum for the S anomaly is more superficial, around 70 m. These subsurface maxima must be associated with the variability of the MAW that is known to display a wide range of T and S characteristics (Benzohra and Millot, 1995), due to different degrees of mixing. This water mass roughly forms a surface layer of about 150 m that is transported eastward along the coast by the Algerian Current. A high mesoscale activity is associated with this current, in the form of meanders and, more frequently, of anticyclonic eddies that strongly influence the distribution of the water masses in the basin and their relative mixing (Millot, 1991).

The first three modes are also the ones that account for the variability at depth. In particular, the second mode displays a coherent structure below 200 m, where T and S show fluctuations of the same sign. A similar structure is displayed by the first mode, which dominates in S. We believe this behavior captures the LIW variability, which is normally subject to compensating effects with respect to temperature and salinity at depths (warm and salty at depth).

The higher modes contribute very little to the signal and, as usual, they represent features of limited vertical extent, as demonstrated by the large number of zero-crossings.

For comparison purposes, the first four modes calculated from the summer data set are shown in Figs. 6c and 6d. The surface variability is the main difference with respect to the winter calculation. It is much greater than in winter and is scattered over the whole spectrum of eigenvalues. In fact, the EOFs actually have large amplitudes at the surface – up to the sixteenth mode (not shown) – although most of the modes are related to noise.

5.1.2 Northwestern Basin

Figure 7a and b show the first six vertical modes for both temperature and salinity calculated from the winter data set. The first mode (40%, see Table 4) contributes to the variance of the whole water column considered here. It contains a great part of the variability of the surface layer and most of the variability of the layer below. The T anomalies have the same sign throughout the entire water column that was considered, while the S anomalies change sign around 180 m. This salinity pattern does not change significantly in summer (the zero crossing in S is just a little deeper than in winter) as one can verify by looking at Figs. 7b and d. The temperature mode instead visibily changes between summer and winter, as seen when Figs. 7a and c are compared. The first T-EOF clearly represents the deep winter mixed layer produced by convective processes, while the S-EOF accounts for the presence below the surface of LIW, and a surface MAW that has a larger signal during the summer due to the stratification conditions.

Mode 2 (21%) contributes to the variance of the first 300–350 m, with an amplitude maximum between 100 and 160 m. This depth indicates the transition from the fresher and colder MAW at the surface and the saltier and warmer LIW below, as shown clearly in Fig. 5b. This subsurface maximum tells us that the variability of this interface is large, especially in winter. During winter, the T/S interface is disrupted due to intense vertical mixing and the EOF show the same sign in the first 350 m. Instead, during summer, the interface between MAW and LIW corresponds to a change in sign of the second EOF at 150 m for temperature (Fig. 7c).

Mode 3 (14%) during winter contributes mainly to the variance of T in the first 200 m, its contribution being equivalent to that of the first mode. During summer, this mode shows a temperature subsurface maximum close to the surface (30 m) which may be indicative of the variability in the summer mixed layer depth. With regard to the salinity, this mode possesses a small amplitude and during winter shows coherent fluctuations from the surface down to 500 m. The higher modes represent smaller vertical scales, approaching the noise level. For instance, mode 4 (8%), in the case of temperature during winter, displays a fluctuation in which the surface and deeper layer co-oscillate, while the intermediate layer oscillates in the opposite direction.

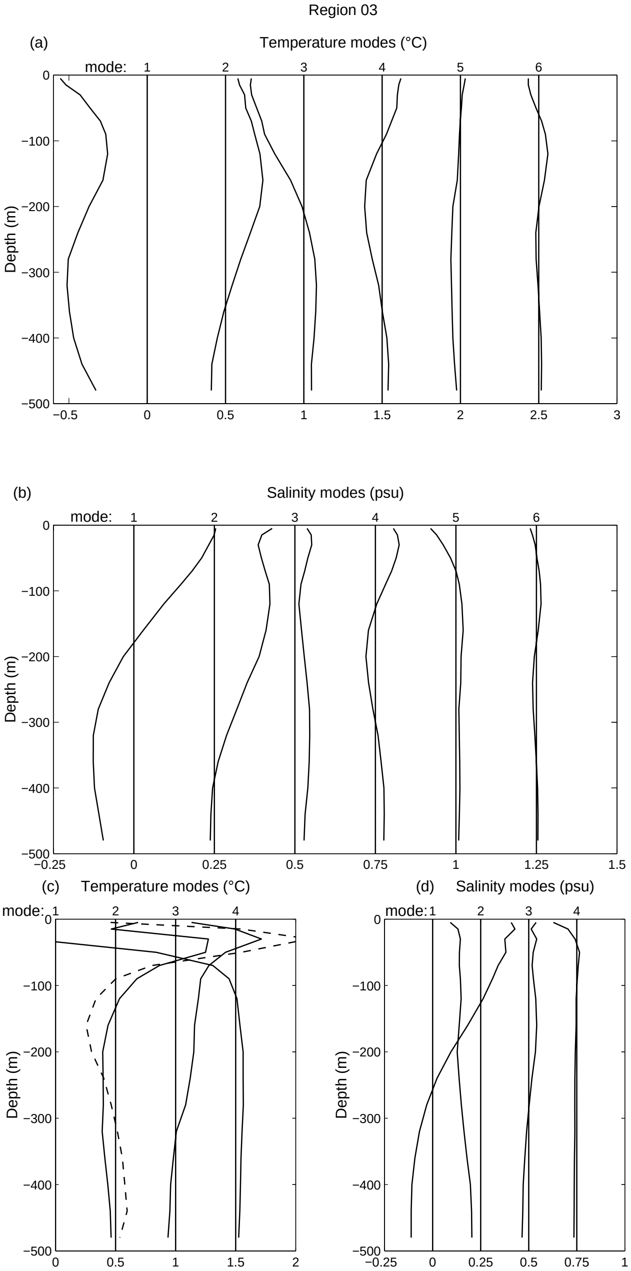


Fig. 7. Vertical EOFs for region 3 (northwestern Basin): **(a)** temperature and **(b)** salinity modes in winter, and **(c)** temperature and **(d)** salinity modes in summer. The first six winter modes and four summer modes are shown. For graphical convenience, a dashed line is used to indicate a change in sign, and respective offsets of 0.5°C and 0.25 psu have been added to each T and S mode for ease of presentation.

Table 5. The first ten eigenvalues of the covariance matrix of the in situ data in region 9 for winter and summer. The first column contains the eigenvalues. The percentage of the variance explained (PVE, Eq. (3) is shown in the second column. In the remaining two columns for each season, the ratio of the vertical average of the root mean square residual (7b) to the vertically averaged standard deviation (7a) is also shown. This latter is shown only for the layer between 0 and 200 m

	Winter				Summer			
Mode	λ	PVE (%)	rms ΔT (°C) 0.458°C	rms ΔS 0.055	λ	PVE (%)	rms ΔT (°C) 0.924°C	rms ΔS 0.095
1	12.7	39.1	0.87	0.88	12.6	38.8	0.78	0.92
2	8.3	25.5	0.81	0.47	8.2	25.3	0.60	0.70
3	6.6	20.4	0.36	0.44	3.1	9.6	0.58	0.55
4	1.4	4.4	0.28	0.34	2.1	6.5	0.50	0.48
5	1.0	3.1	0.27	0.26	1.6	5.1	0.41	0.45
6	0.9	2.7	0.21	0.22	1.3	4.1	0.33	0.37
7	0.38	1.2	0.18	0.19	0.7	2.3	0.30	0.32
8	0.30	0.9	0.15	0.16	0.49	1.5	0.26	0.28
9	0.18	0.6	0.13	0.15	0.43	1.3	0.23	0.26
10	0.14	0.4	0.12	0.15	0.37	1.2	0.21	0.21

5.1.3 Levantine Basin

Figures 8a and b show the first six vertical modes for the winter. Most of the temperature variability is contained in the first three modes, as evident in Fig. 8a, where we had to use different horizontal scales to distinguish the first from the last three temperature modes. The first mode (39%, see Table 5) indicates coherent fluctuations from the surface down to 500 m for both T and S that furthermore have the same sign. A subsurface maximum is found at approximately 300 m for temperature. This behavior can be attributed to the well-mixed winter vertical structure of S in the Levantine (Fig. 5c) which also forces the vertical coherence in the first T-EOF.

The second mode (24%) evidences two vertical structures, the first above 250 m, where both the parameters are subject to big variations, and the second below 250 m, where only T shows significant variations. Around this depth, an inflection is observed in the mean T profile (Fig. 5c), marking the separation between saltier and warmer surface waters and the fresher (by just 0.1 psu) and colder water of the intermediate layer. Above 250 m, the second temperature EOF shows a zero crossing around 100 m. In the portion of the water column extending from the surface down to 100 m, S and T fluctuate with the same sign. Below, they are opposite in sign. Interpreting this EOF as representative of the variability of such T-S structure (same sign anomalies at the surface and opposite signs below 100 m), we think this is indicative of eddy-like features with a high baroclinic vertical structure. Mode 3 (20%) represents mainly the T variations above 300 m.

The contribution of the higher modes is much less than in the other regions. This makes us conclude that in the Levantine region, we need less vertical EOFs than in the western basin to represent the T-S variability structure.

The first four modes calculated from the summer data set are shown in Figs. 8c and d. As for the other regions, a large part of the variability is observed in T in the first 100 m (maximum amplitude around 50 m). This is the variability associated with the formation of the summer warm mixed layer. Below 100 m, the LIW structure is reconstructed by the first two salinity modes that now have larger amplitudes than during winter. Thus, the summer period in the Levantine region is characterized by large variability in the vertical structure of the salinity anomalies, irrespective of whether they are of the same or opposite sign with respect to the sign of the corresponding temperature anomalies.

5.2 Results from the model simulations

The first three modes obtained from the model data in winter are shown in Fig. 9. Even though we cannot expect them to be exactly similar to the modes computed from the in situ data, we will show here that the model simulation is capable of recovering some of the major features of the in situ EOFs.

In region 2 (Figs. 9a and b), we found that most of the variability is in the upper 200 m, and its signal is shared among the modes represented, as was also observed while analyzing the results from the in situ data. But differences are found in the amplitude of the temperature EOFs which is twice that of the corresponding EOFs obtained with the in situ data. The salinity EOFs are instead quite similar in amplitude and shape in the case of both the data sets (compare Figs. 6b with 9b). In region 3 (Figs. 9c and d), the temperature EOFs again show an amplitude that is twice that of EOFs computed from the in situ data set (Fig. 7a), while the amplitudes in the case of the salinity are similar. The shape of the temperature EOF reveals some similarity to the observed EOF, but in general

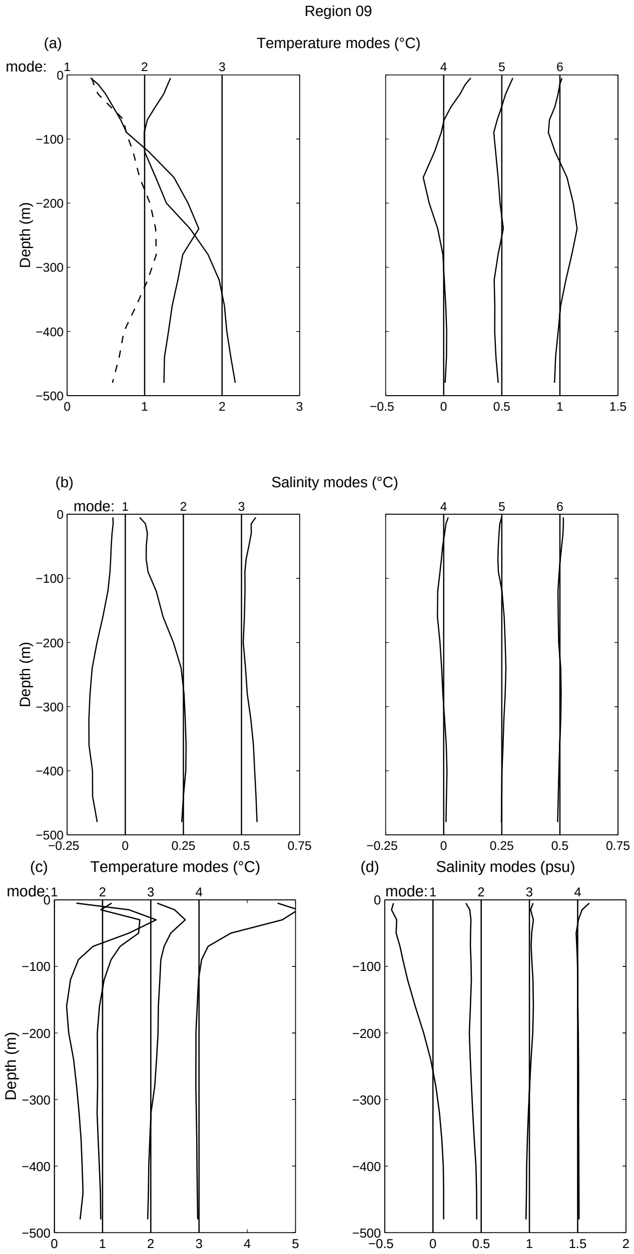


Fig. 8. Vertical EOFs for region 9 (Levantine Basin): **(a)** temperature and **(b)** salinity modes in winter, and **(c)** temperature and **(d)** salinity modes in summer. The first six winter modes and four summer modes are shown. For graphical convenience, a dashed line is used to indicate a change in sign. For ease of presentation, offsets of 1°C and 0.5°C have been added to the first and the second triplets of T modes, respectively, and an offset of 0.25 psu has been added to each S mode.

Table 6. The order of the optimum reduced set of EOFs as indicated by a value of the ratio of the root mean square residuals to the standard deviations of 30%. The percentage of total variance explained is shown in parentheses

Region	Win	Spr	Sum	Aut	Year
Alboran	7 (94%)	8 (94%)	8 (94%)	9 (95%)	8 (93%)
Algerian	8 (94%)	8 (93%)	9 (94%)	9 (94%)	9 (93%)
North-Western	7 (93%)	9 (93%)	10 (95%)	8 (93%)	9 (93%)
Tyrrhenian	7 (93%)	6 (92%)	8 (92%)	9 (94%)	9 (93%)
Sicily Strait	7 (95%)	-	8 (94%)	7 (93%)	9 (93%)
Ionian	5 (93%)	7 (94%)	8 (93%)	8 (94%)	8 (93%)
Adriatic	–	–	–	–	6 (95%)
Aegean	3 (90%)	–	–	–	7 (94%)
Levantine	5 (92%)	7 (93%)	8 (93%)	8 (94%)	9 (94%)

for region 3, there is no evidence of very deep mixing, as was instead evident in the case of the observed EOF. For region 9 (Figs. 9e and f), the temperature EOFs show the same amplitude as the observed EOF (Fig. 8a), but they are more surface intensified, again showing a difference between observed and model data in the depth range of mixing processes. In all the regions, the temperature and salinity EOFs show the same relationships as was found in the EOFs calculated from the in situ data, which means that the overall T-S characteristics obtained with the model data agree with those derived from the observations.

To better evaluate the difference between the EOFs computed from the model and the in situ data, we will use the former as a basis to reconstruct in situ profiles. At the same time, we reconstruct the in situ profile using the EOFs derived from the same in situ data, clearly the "best" that we can do. In other words, we use in Eq. (5) alternatively the $\boldsymbol{e}_i$ from the in situ and the model data. The square root of Eq. (6) is then used to evaluate the rms residual error for the two cases. This projection method is mimicking what an assimilation system, such as SOFA, would do using the order reduction EOFs from model data to assimilate temperature profiles. We specify that the projection is done on the full set of EOFs, and the reconstruction is done using a reduced set. This calculation is shown for region 3 in Fig. 10 and for region 9 in Fig. 11 for the winter season.

The rms residual errors for T and S are shown for $k = 2, 4, 6, 8$, and 10, where k is the number of retained modes. From the comparison, we deduce that the EOFs calculated from the model data are as good as the EOFs deduced from the observational data set to reproduce the in situ profiles. The rms residual error when using EOFs from model data is even slightly less than the corresponding residual error obtained by using EOFs from in situ data. This is mainly due to the homogeneous sampling of model data both in space and time against the shortage of the in situ data.

In region 9, the rms residual error for temperature and salinity using model EOFs is actually smaller than the corresponding error obtained with in situ EOFs. We believe this can be explained by the scarcity of in situ data used in the calculation of T, S anomalies for this region (see Table 1) and the uncertainty in the climatology.

In conclusion, we could say that both model simulation EOFs and in situ EOFs are capable of reproducing the essen-

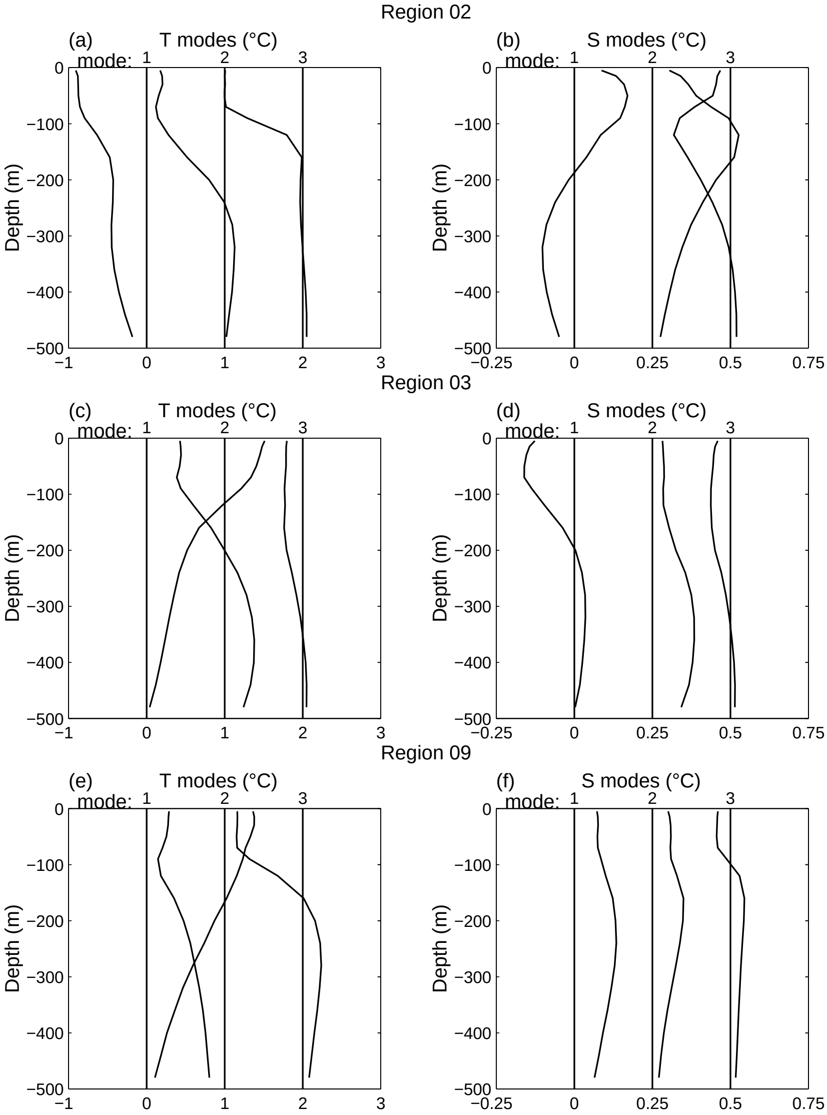


Fig. 9. Winter EOFs from model data. Temperature **(a)** and salinity **(b)** for the Algerian Basin, temperature **(c)** and salinity **(d)** for the northwestern Basin, and temperature **(e)** and salinity **(f)** in the Levantine Basin. The first three modes are shown.

tial thermohaline variability of the Mediterranean Sea historical data. The details of the structure of each mode differ between the model and in situ EOFs, but the information content of the two EOF sets is equivalent with respect to their ability to reproduce the observed variability.

6 The reduced order EOF space

In this section we evaluate the optimal number of EOFs to be used in reconstructing the full data signal, and their statistical significance. As already explained, to identify the significant and optimal EOF subspace, we can use a number of indices. In Tables 3, 4 and 5, we show the calculation for the PVE and the ratio between Eq. (7a) and Eq. (7b) in the first 200 m of the water column. This is the portion of the water column where the variability is the highest. As already anticipated in Sect. 3.2, we will define as optimal the set of modes that return cumulatively a root mean square residual ratio of less than 30%, a threshold value which ensures that a variance greater than 90% is explained.

In Table 3, we show the calculations for region 2 and the winter and summer seasons. We conclude that eight modes in winter and nine in summer are enough to explain the bulk of the variability. With this selection, we explain 94% of the total variance. The first 3 modes explain almost 70% of the variance, and the first mode by itself explains 32–36% of the variance.

Table 4 summarizes the same results for region 3. The root mean square residual error ratio indicates as dominant the first seven EOFs in winter and the first ten in summer, which explain, respectively, 93% and 95% of the total variance. As before, most of the variance (72–76%) is represented by the first three modes.

The results for region 9 are shown in Table 5. The root mean square residual error ratio indicates five dominant EOFs in winter and eight in summer, explaining 92–93% of the total variance. Once again, the first three modes account

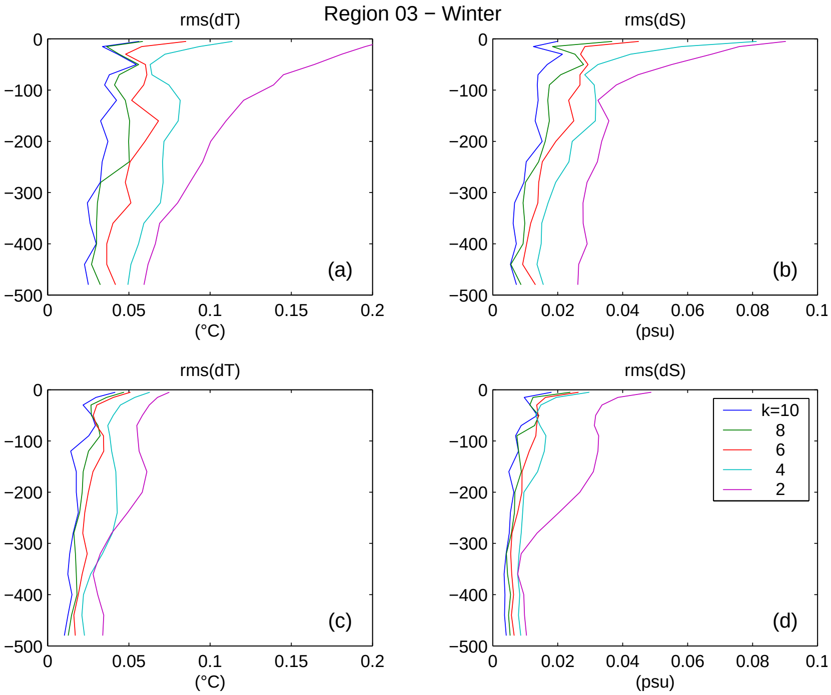


Fig. 10. RMS residual errors of temperature and salinity for region 3 in winter obtained by projecting the in situ data on a k-order EOF subspace, with $k = 2, 4, 6, 8$ and 10 when EOFs from in situ data (**a** and **b**) and model data (**c** and **d**) are used.

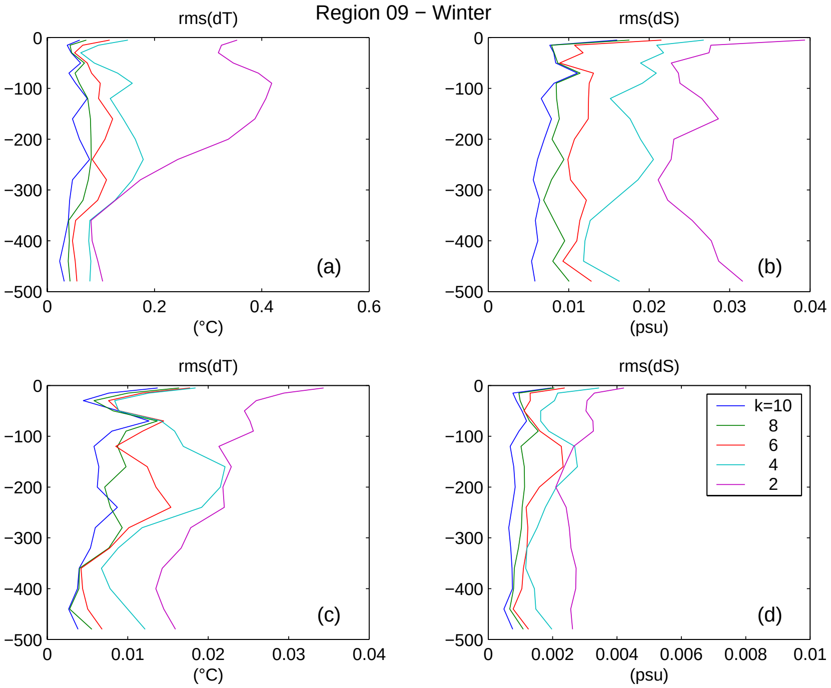


Fig. 11. RMS residual errors of temperature and salinity for region 9 in winter obtained by projecting the in situ data on a k-order EOF subspace, with $k = 2, 4, 6, 8$ and 10 when EOFs from in situ data (**a** and **b**) and model data (**c** and **d**) are used.

for most of the variance that, in this region, is greater in winter (85%) than in summer (74%).

To empirically infer the statistical significance of each EOF, in Fig. 12, we compare the spacing of the first ten eigenvalues, λ_k, with their sampling error, $\delta\lambda_k$, as a function of the mode number k. Only in region 3 are the data sufficient to ensure that each EOF is properly resolved from the adjacent ones. In the other regions, the data are not enough, and some degeneracy is found. For example, in region 2, in winter, the sampling errors of the first and second eigenvalues, and of the fourth and fifth eigenvalues are comparable to their spacings, consequently, the corresponding EOFs are not statistically different. Other multiplets are observed at higher mode numbers, placed in the noisy part of the eigenvalue spectrum. The same occurs in the winter calculation for region 9, where the second and third eigenvalues are so close that the error bars overlap. As explained in Sect. 3.3, when degeneracy occurs we cannot be sure of having found the real “shape” of a

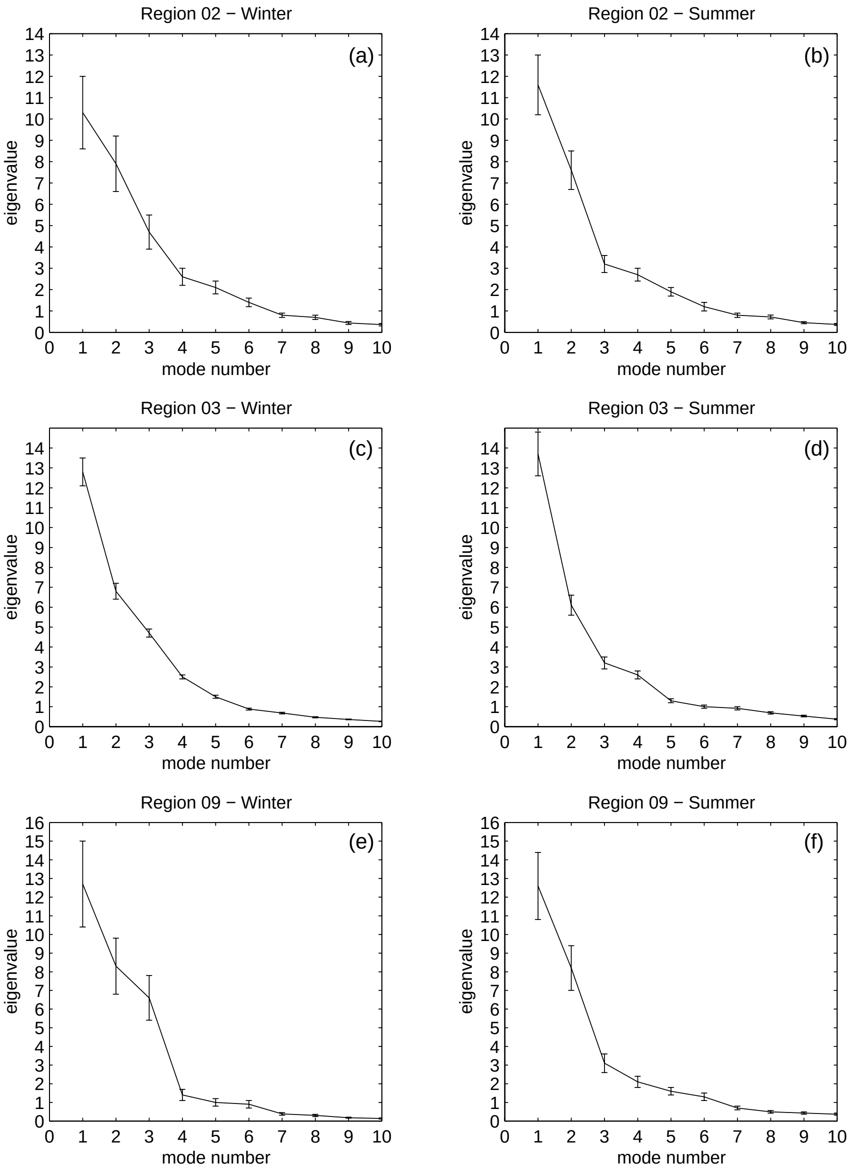


Fig. 12. Diagram of the first ten eigenvalues (λ_k) of the in situ EOFs versus the mode number k in winter **(a)** and summer **(b)** for the Algerian Basin, winter **(c)** and summer **(d)** for the northwestern Basin, and winter **(e)** and summer **(f)** in the Levantine Basin. The error bars represent the sampling errors ($\delta\lambda_k$).

degenerate EOF, since any linear combination of degenerate EOFs can also be an eigenvector.

In Table 6, we show the optimal and reduced number of EOF modes calculated from in situ data for all the regions and all the seasons. We observe that, in winter, the number of retained modes is less than in the other seasons. A difference also exists between the western and the eastern regions. Independently of the season, the variability existing in the eastern Mediterranean is described by a smaller number of modes. As a word of caution, we would like to point out that generally the amount of data is less in the eastern than in the western regions, so this result could be simply due to a subsampling of the variability.

The variance explained by the optimal set of modes is 90–95%. This is really a high percentage, and one can object that we were too conservative in setting the rms residual error ratio to 0.3. Depending on the particular scientific problem that the EOF subspace is used for, some physical considerations can help to further reduce the order. When it is used to assimilate altimeter data, for instance, a further criterion can be used that enforces the condition that the information contained in a T-S profile is also observable by the altimeter, i.e. the T-S signal is present in the sea level anomaly. Faucher et al. (2002), for instance, evaluate the contribution of each mode to the dynamic height anomaly (DHA), an approximation of the sea level anomaly at mid-latitudes and for the open ocean, by calculating the ratio between the root mean square of the unresolved DHA and the root mean square of the total DHA. They accept a ratio of 0.4. When this criterion is applied to our data, the optimal number of modes decreases drastically to 1–3, depending on the region/season, and it remains small when the threshold is decreased from 0.4 to 0.1–0.2.

7 Summary and conclusions

In this paper, we have attempted to provide a representation of the upper thermohaline vertical structure of the Mediterranean Sea by means of bivariate vertical EOFs, using temperature and salinity from both observations and model simulations. The focus of this work was twofold: first, to contribute to the description of the Mediterranean Sea thermohaline structure for general purposes and second, to find the optimal number of vertical modes that can reproduce the T-S characteristics. This knowledge is the basis of the correct assimilation of in situ data by means of a multivariate Optimal Interpolation scheme used in MFSPP (Demirov et al., 2003).

Extending the analysis from the surface to 480 m, it was shown that most of the thermohaline variability is associated with the surface layer, usually occupied by MAW, even if a smaller signal is evident at the intermediate levels as well, where LIW is found. Most of the variability observed in the surface layer has to be ascribed to seasonal and mesoscale processes. Generally, the first three modes are able to reproduce most of the variance contained in the data; their dominance over the others is remarkably high in the eastern basins. The higher modes contribute very little to the signal and represent smaller vertical scales approaching the noise level.

The calculations from model simulations and observations were not in perfect agreement. In fact, the former generally overestimates the amplitudes of the temperature EOFs with respect to the corresponding ones obtained from the in situ data, besides sometimes reproducing different features. However, we have demonstrated that both sets of EOFs, those from the model simulations and the observations, have the same ability to reproduce the essential variability present in the observed data. This result is very important from the point of view of assimilation, since the EOFs from model simulations may be thought of as an alternative set of basis functions upon which to project the thermohaline variability of a particular data set.

We paid particular attention to the study of the "significant" bivariate EOFs and the "optimal" reduced set. To evaluate the optimal number of EOFs that can reconstruct the full data signal, we calculated standard indices such as the Percentage of Variance Explained by each mode and the root mean square residual error. The latter was compared as a ratio to the observed averaged standard deviation of each variable, to ensure that the residual of the projection was sufficiently small with respect to the data variability. We found that the number of EOFs needed to capture the variability contained in the original data changes with geographical region and season. In particular, winter data require a smaller number of modes (4–8, depending on the region) than the other seasons (8–9 in summer). Moreover, independently of the season, the variability existing in the eastern Mediterranean is described by a smaller number of modes than in the western Mediterranean. We cannot exclude that this result could be simply due to a subsampling of the variability, associated with the data scarcity in the eastern regions.

To summarize, we have shown that the large-scale Mediterranean thermohaline vertical structure can be represented by a limited number of vertical bivariate EOFs, and that the "optimal set" can be selected on the basis of general principles. Future calculations that will include in the state vector other variables such as the stream function and the sea surface pressure are planned, in order to take into account the barotropic part of the transport that is overlooked when the vertical modes are calculated from a hydrographical database only.

Acknowledgements. We would like to thank Dr. A. Navarra and Dr. P. De Mey for interesting discussions and comments during the work. Special thanks to P. Faucher for his useful suggestions regarding the EOF methodology. Finally, we wish to thank Mr. R. Nair for the revision of the English text. This research was conducted under the MFS Pilot Project contract EU MA53-CT98-0171.

The Editor in chief thanks C. Sammari, S. Marullo and another Referee for their help in evaluation this paper.

References

Benzohra, M. and Millot, C.: Characteristics and circulation of the surface and intermediate water masses off Algeria, Deep Sea Res., 42(10), 1803–1830, 1995.

Bignami, F., Marullo, S., Santoleri, R., and Schiano, M. E.: Long wave radiation budget in the Mediterranean Sea, J. Geophys. Res, 100, 2501–2514, 1995.

Brankart, J. M. and Pinardi, N. : Abrupt cooling of the Mediterranean Levantine Intermediate Water at the beginning of the eighties: observational evidence and model simulation, J. Phys. Oceanogr., 31, 8(2), 2307–2320, 2001.

Brasseur, P., Beckers, J. M., Brankart, J. M., and Schoenauen, R.: Seasonal Temperature and Salinity Fields in the Mediterranean Sea: Climatological Analyses of a Historical Data Set, Deep Sea Res., 43(2), 159–192, 1996.

Castellari, S., Pinardi, N., and Leaman, K. D.: A model study of airsea interactions in the Mediterranean Sea, J. Mar. Syst., 18, 89–114, 1998.

Castellari, S., Pinardi, N., and Leaman, K. D.: Simulation of water mass formation processes in the Mediterranean Sea: influence of the time frequency of the atmospheric forcing, J. Geophys. Res., 105 (N 10), 24 157–24 181, 2000.

Cox, M.: A primitive equation, 3-dimensional model of the ocean, GFDL Ocean Goup Tech. Rep., vol. 1., Geophys. Fluid Dyn. Lab., Princeton, NJ, 43pp, 1984.

Da Silva, A. M., Young, Ch., and Levitus, S.: Atlas of surface marine data 1994, NOAA Atlas NESDIS 7, 1994.

De Mey, P.: Data assimilation at the oceanic mesoscale: a review, J. Met. Soc. Japan, Special issue on "Data assimilation in meteorology and oceanography: Theory and practice", 75, 415–425, 1997.

De Mey, P., and Benkiran, M.: A multivariate reduced-order optimal interpolation method and its application to the Mediterranean basin-scale circulation, in: Pinardi, N. and Woods, J. D. (Eds.): Ocean Forecasting: Conceptual basis and applications, Springer-Verlag, 281–306, 2002.

De Mey, P. and Robinson, A. R.: Assimilation of altimeter eddy fields in a limited-area quasi-geostrophic model, J. Phys. Oceanogr., 17, 2280–2293, 1987.

Demirov, E. and Pinardi, N.: Simulation of the Mediterranean Sea circulation from 1979 to 1993. Part I: The interannual variability, J. Mar. Syst., 33/34, 25–30, 2002.

Demirov, E., Pinardi, N., De Mey, P., Tonani, M., Fratianni, C., and Giacomelli, L.: Assimilation scheme of the Mediterranean Forecasting System: Operational Implementation, Ann. Geophysicae, this issue, 2003.

Faucher, P., Gavart, M., and De Mey, P.: Isopycnal EOFs in the North and Tropical Atlantic and their use in estimation problems, accepted, J. Geophys. Res., 2002.

Fichaut, M., Balopoulos, E., Dooley, H., Garcia-Fernandez, M. J., Iona, A., Jourdan, D., Baudet, L., and Maillard, C.: A common protocol to assemble a coherent database from distributed heterogeneous data sets: The MEDATLAS database experience, in: Marine science and technology programme: Experiences in project data management, Office for Official Publications of the European Communities, Luxembourg, 311–327, 1998.

Frankignoul, C. and Reynolds, R.: Testing a Dynamical Model for Mid-Latitude Sea Surface Temperature Anomalies, J. Phys. Oceanogr., 13, 1131–1145, 1983.

Fukumori, I. and Wunsch, C.: Efficient representation of the North Atlantic hydrographic and chemical distributions, Prog. Oceanogr., 27, 111–195, 1991.

Gavart, M. and De Mey, P.: Isopycnal EOFs in the Azores Current Region: A Statistical Tool for Dynamical Analysis and Data Assimilation, J. Phys. Oceanogr., 27, 2146–2157, 1997.

Kondo, J.: Airsea bulk transfer coefficients in diabatic conditions. Boundary Layer Meteorol., 9, 91–112, 1975.

Korres, G., Pinardi, N., and Lascaratos, A.: The Ocean Response to Low-Frequency Interannual Atmospheric Variability in the Mediterranean Sea. Part I: Sensitivity Experiments and Energy Analysis, J. Climate, 13, 705–731, 2000a.

Korres, G., Pinardi, N., and Lascaratos, A.: The Ocean Response to Low-Frequency Interannual Atmospheric Variability in the Mediterranean Sea. Part II: Empirical Orthogonal Functions Analysis, J. Climate, 13, 732–745, 2000b.

Hecht, A., Pinardi, N., and Robinson, A. R.: Currents, Water Masses, Eddies and Jets in the Mediterranean Levantine Basin, J. Phys. Oceanogr., 18, 10, pp. 1320–1353, 1988.

Hellerman, S. and Rosenstein, M.: Normal monthly wind stress over the world ocean with error estimates, J. Phys. Oceanogr., 23, 1009–1039, 1983.

Levitus, S., Gelfeld, R., Boyer, T., and Johnson, D.: Results of the NODC Oceanographic Data and Archaeology and Rescue Project. Key to Oceanographic Records Documentation; 19–73, 1994.

Lorenz, E. N.: Empirical Orthogonal Functions and Statistical Weather Prediction, Statistical Forecasting Project – Scientific Report No. 1, Dept. Of Meteorology, Massachusetts Institute of Technology, 49pp, 1956.

Maes, C.: A note on the vertical scales of temperature and salinity and their signature in dynamic height in the western Pacific Ocean: Implication for data assimilation, J. Geophys. Res., 104 (C5), 11 037–11048, 1999.

Millot, C.: Mesoscale and seasonal variabilities of the circulation in the western Mediterranean, Dyn. Atm. Oceans, 15, 179–214, 1991.

Nittis, K., Pinardi, N., and Lascaratos, A.: Characteristics of the summer 1987 flow field in the Ionian Sea, J. Geophys. Res., 98, C6, 10 171–10 184, 1993.

North, G. R., Bell, T. L., Cahalan, R. F., and Moeng, F. J.: Sampling errors in the estimation of empirical orthogonal functions, Mon. Weather Rev., 110, 699–706, 1982.

Pinardi, N. and Masetti, E.: Variability of the large-scale general circulation of the Mediterranean Sea from observations and modelling: a review, Palaeogeography, Palaeoclimatology, Palaeoecology, 158, 153–173, 2000.

Pinardi, N., Allen, I., Demirov, E., De Mey, P., Korres, G., Lascaratos, A., Le Traon, P.-Y., Maillard, C., Manzella, G., and Tziavos, C.: The Mediterranean ocean Forecasting System: first phase of implementation (1998–2001), Ann. Geophysicae, this issue, 2003.

Pinardi, N., Auclair, F., Cesarini, C., Demirov, E., Fonda-Umani, S., Giani, M., Montanari, G., Oddo, P., Tonani, M., and Zavatarelli, M.: Toward marine environmental predictions in the Mediterranean Sea coastal areas: a monitoring approach, in: Ocean Forecasting: Conceptual basis and applications, (Eds) Pinardi, N. and Woods, J. D., Springer-Verlag, 339–376, 2002.

Preisendorfer, R. W., Zwiers, F. W., and Barnett, T. P.: Foundations of principal component selection rules. SIO Ref. Ser. 81-4, Scripps Institution of Oceanography, pp. 192, 1981.

Preisendorfer, R. W.: Principal Component Analysis in Meteorology and Oceanography, (Ed) Mobley, C. D., Elsevier, Amsterdam, pp. 452, 1988.

Reed, R. K.: On estimating insolation over the ocean, Progr. Oceanogr, 17, 854–871, 1977.

Roussenov, V., Stanev, E., Artale, V., and Pinardi, N.: A seasonal model of the Mediterranean Sea circulation, J. Geophys. Res., 100, 13 515–13 538, 1995.

Von Storch, H. and Hannoschöck, G.: Statistical Aspects of Estimated Principal Vectors (EOFs) Based on Small Sample Size, J. Clim. Appl. Meteorol., 24, 716–724, 1986.