Reconstructing the Ocean State Using Argo Data and a Data-Driven Method

Erwan Oulhen, Nicolas Kolodziejczyk, Pierre Tandeo, Bruno Blanke, Florian Sévellec

Reconstructing the Ocean State Using Argo Data and a Data-Driven Method

Erwan Oulhen,[a] Nicolas Kolodziejczyk,[a] Pierre Tandeo,[b,c,d] Bruno Blanke,[a] Florian Sévellec,[a,c]

[a] *Univ Brest, CNRS, IRD, Ifremer Laboratoire d'Océanographie Physique et Spatiale (LOPS, UMR 6523), IUEM, Brest, France*

[b] *IMT Atlantique, Lab-STICC, UMR CNRS 6285, 29238, Brest, France*

[c] *Odyssey, Inria/IMT/CNRS, Rennes, France*

[d] *RIKEN Center for Computational Science, Kobe, 650-0047, Japan*

Corresponding author: erwan.oulhen@univ-brest.fr

ABSTRACT: Twenty years of Argo profiles have provided unprecedented insight into the global space and time variability patterns of ocean temperature and salinity, significantly reducing associated uncertainties. However, such assessments during the pre-Argo period remain a challenge due to the scarcity of observations in many regions. From the Argo period, a set of dominant three-dimensional patterns can be estimated using EOF analysis and used to fill in observational gaps. From the associated principal components, temporal fluctuations can be observed, aiming to build a catalog of possible trajectories of the ocean state. To map pre-Argo observations, EOFs are used in a data assimilation framework using a catalog to feed an analog prediction and provide reanalysis. In this study, this new data-driven interpolation method is called RedAnDA (Reduced-space Analog Data Assimilation) and was tested in the tropical Pacific Ocean. RedAnDA was first validated through an Observing System Simulation Experiment (OSSE) approach, using synthetic observations extracted from a model simulation. It was then applied to a real historical dataset and compared to other available reanalysis products. Overall the reconstructed temperature field showed variability consistent with the OSSE application and other reanalysis products in the real data application. Further improvements are needed to optimally estimate uncertainty, but RedAnDA already combines valuable information about state predictability, observation sampling, and unresolved scale issues.

1. Introduction

In the ocean, temperature, by setting the density in most of tropical to subpolar regions, plays a fundamental role in influencing dynamical processes such as three-dimensional geostrophic circulation and the mixing of water masses. Temperature has a first order impact on ecosystems (**?**). The globally averaged Ocean Temperature temporal changes reflect long-term global warming through heat uptake (**??**), but also climate and atmospheric variability by air-sea interaction (**?**). Moreover, the ocean exhibits significant variability spanning regional to basin scales and ranging from seconds to decades, which can superimpose and interact, increasing uncertainties in estimating long-term anthropogenic trends or underlying natural variability (**???**). Notably, phenomena such as the air-sea coupled El Niño-Southern Oscillation (ENSO) have a pronounced impact on surface temperature primarily in the Pacific Ocean, with global impacts on long-term temperature trends (**??**). Consequently, evaluating the influence of natural oceanic temperature variability represents a crucial challenge. Therefore, precise estimates of global and regional oceanic variability derived from *in situ* measurements remain timely.

From the late 1930s to the 1990s, successive advances in *in situ* measurement technology facilitated the expansion of observational sampling of the ocean, particularly along repeated hydrographic transects in the major ocean basins and sub-basins. However, a notable bias persisted in the distribution of observations, in favor of a better sampling the Northern Hemisphere and coastal regions (**?**). Furthermore, the majority of research expeditions were conducted during summer months, primarily due to favorable weather conditions for navigation. This lack of *in situ* observations has impeded our ability to consistently describe regional patterns and understand underlying oceanic mechanisms (**?**), thereby contributing to uncertainties in quantifying the rate of ocean heat content change (**?????**).

Since the late 1990s, the Argo global observing system has been deployed, achieving in 2007, its targeted quasi-global coverage of 3x3° T and S profiles over the upper 2000 m of the ocean (**??**). Over the last two decades, it has become the backbone of the Global Ocean Observing System (GOOS) and is now the primary source of observation for numerous ocean and climate studies (see **?**, **?**, for a recent review). The Argo system effectively addresses the major sampling biases inherent in ocean observing systems, reducing uncertainties in global oceanic records and

enhancing the confidence in assessments of ocean heat content, thermal expansion, and changes in freshwater content (Hansen et al. 2011; **?**; **?**; **?**).

One of the principal challenges in oceanography lies in filling the gaps in sparsely distributed observations, especially during the pre-Argo period, i.e., before 2000. Monitoring of the global ocean state is significantly enhanced when observational variables are remapped on a regular grid utilizing statistical methodologies or assimilated into a numerical model. Among the procedures for mapping oceanic data, several methods can be categorized.

Firstly, widely used mapping approaches are methods that rely solely on observations and *a priori* statistics, such as Optimal Interpolation (OI, Bretherton et al. 1976). Nevertheless, they exhibit limited skill in accurately estimating error. Among the notable Optimal Interpolation applications, Good et al. (2013) reconstructed past temperature and salinity variability, from 1900 onwards, with a focus on evaluating observational quality. Kaplan et al. (1997) combined data reduction using Empirical Orthogonal Functions and least squares mapping techniques (e.g., optimal smoother, Kalman filter, and OI) to analyze Sea Surface Temperature (SST) anomalies in the Atlantic Ocean for the period 1856-1991. This approach was computationally efficient and yielded consistent results with robust error estimates. A last example is **?**, who utilized an ensemble OI approach, incorporating a first guess and *a priori* covariance derived from the 5$^{\text{th}}$ phase of the Coupled Model Intercomparison Project (CMIP5), to reconstruct ocean subsurface temperature patterns and the trend from 1940 to 2014.

Secondly, data assimilation procedures in numerical models allow for a physically consistent data interpolation (**?**). These analyses give robust estimates of upper Ocean Heat Content (OHC) and the depth of the tropical mixed layer (**?**). However, this approach requires significant numerical costs, as it requires the integration of the primitive equations into an Ocean General Circulation Model (OGCM) and are subject to inherent numerical imperfections (**?**).

In recent years, a new class of methods has emerged, taking advantage of the ever-increasing amount of data available from both satellite and *in situ* measurements, and recent advances in applied mathematics. These data-driven methods have been specifically developed for data assimilation and climate prediction purposes (e.g., **?????**). Leveraging Argo data, **?** employed stationary Gaussian process regression to compute temperature fields, estimating covariance parameters through local moving-window maximum-likelihood estimation. They have enhanced uncertainty

estimates by recognizing the non-Gaussian and non-stationary nature of the temperature field. **?** introduced PROCAST, an efficient system that probabilistically predicts global mean air and sea surface temperature variations, utilizing insights from CMIP5 simulations. Notably, data-driven methods offer the advantage of providing physically meaningful error estimates and satisfactory probabilistic forecasts. Additionally, they entail relatively lower numerical costs compared with OGCMs.

In this current study, the Analog forecasting method (**?**) was combined with ensemble Data Assimilation techniques, forming the AnDA method, as previously introduced by **?** and Lguensat et al. (2017). This method has already shown promising results in previous oceanic applications, such as the successful reconstruction of the sea surface height in the Gulf of Mexico (Zhen et al. 2020). What distinguishes this approach from traditional data assimilation schemes is its integration of analog forecasting. By statistically learning the dynamics of the system from a comprehensive catalog of analogs - encompassing a wide range of potential system states along with their subsequent states - the method is adept at forecasting any given state. Consequently, it can confidently estimate the evolution of a state at a given time toward the subsequent time step when the conditions allow predictability, or, alternatively, remain uncertain and stagnant when the state is less predictable.

The aim of the study was to acquire knowledge by constructing a catalog of analogs from a gridded analysis over the well-sampled Argo period and to train AnDA to extrapolate temperature observations over a less well-sampled period (prior to Argo). This approach enabled the assessment of new information on temperature variability, both at the surface and at depth, during a period when *in situ* measurements were lacking. To meet the challenge of processing observations that are sparse in time and three-dimensional in space, such as a large set of hydrographic profiles, the method was combined with the Reduced-Space Interpolation introduced by Kaplan et al. (1997). Specifically, the parameters interpolated by AnDA are the temporal coefficients of the spatial-temporal signals projected into a reduced set of three-dimensional Empirical Orthogonal Functions (EOFs), which are built using the learning period. Utilizing EOFs offers the advantage of extracting dominant patterns of variability, smoothing the analyzed field, and reducing the dimensionality of the system. **?** have demonstrated the effectiveness of using EOFs for reconstruction purposes in the Pacific. The integration of space reduction and AnDA is termed RedAnDA.

The test region selected for implementing the ocean temperature reconstruction was the tropical Pacific. This basin exhibits subsurface OHC variability, predominantly driven by the interannual to decadal ENSO climate mode (e.g., **?**). The ENSO phases redistribute heat throughout the basin, with El Niño events resulting in warming of the upper 100 meters and cooling of the subsequent 400 meters, whereas La Niña events exhibit about the reverse pattern. This well-documented interannual variability has been extensively studied (e.g., **???**). During the Argo period, i.e., the learning period for the catalog, 6 El Niño and 5 La Niña events were observed.

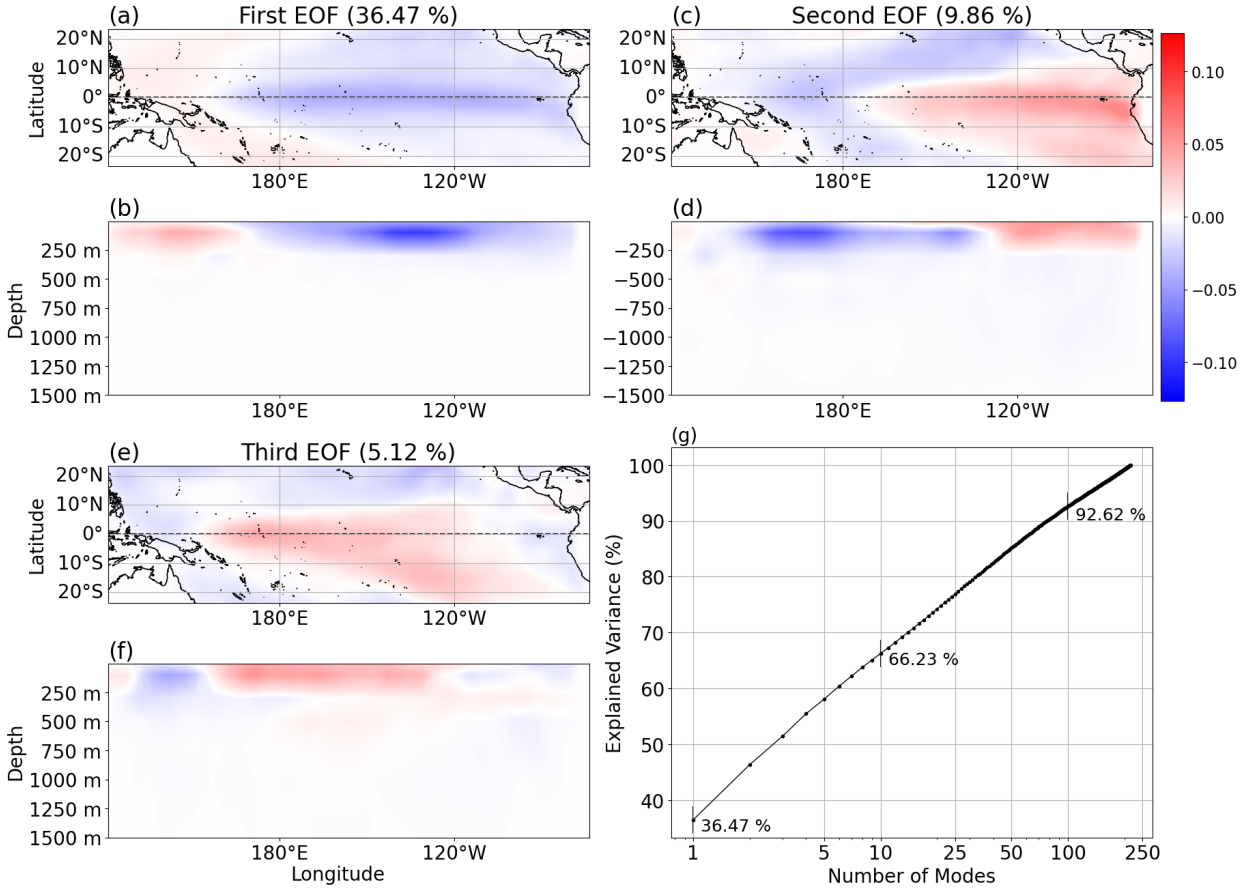
The paper is organized as follows. Firstly, the various procedures of the RedAnDA method are detailed. The third section presents the different datasets used in this study. The fourth section presents the validation of the method using an Observing System Simulation Experiment (OSSE) approach. Finally, the application of RedAnDA to real observations and its comparison are analyzed.

2. The RedAnDA Method

The RedAnDA method is a combination of three established methodologies: i) Space Reduction utilizing Empirical Orthogonal Functions, ii) Analog Prediction, and iii) Data Assimilation.

After removing the monthly seasonal cycle, the three-dimensional temperature anomaly field $\mathbf{T}(t)$ from the learning period, in the tropical Pacific Ocean, is considered. Empirical Orthogonal Function (EOF) decomposition is then applied to reduce the dimension of the temperature field. This method concentrates the variance into a reduced number of functions and relegates insignificant portions to noise (Emery and Thomson 2001). This space reduction facilitates the Analog Prediction, as discussed later in this section. By retaining the L dominant EOFs, the noise is efficiently filtered out at the expense of a relatively small truncation error $\boldsymbol{\epsilon}^r$. The selection of L involves a trade-off between system reduction and variance representation. Before the decomposition, the prior weighting $\mathbf{w} \times \mathbf{T}(t)$, where $\mathbf{w}$ represents the square root of the cosine of latitude, ensures a uniform impact in terms of variance across different regions, accounting for their surface area (Baldwin et al. 2009; Shea 2013).

The respective temporal coefficients of the EOFs, denoted $\boldsymbol{\alpha}(t)$ set, are obtained by projecting the spatial-temporal signals onto the set of eigenvectors. Variations in $\boldsymbol{\alpha}(t)$ indicate changes in the covariance patterns over time. Thus, the field approximation can be expressed as follows:



142 FIG. 1. First (a,c), second (b,d), and third (e,f) EOF computed from ISAS20 temperature at 5-meter depth and
143 along an equatorial section. The explained variance for each of the three EOFs are given in the title of panels
144 (a), (b), and (e). (g) Cumulative explained variance for 252 modes, as the number of EOFs was increased, with
145 the x-axis presented on a log-scale. The values of the explained variance for 1, 10, and 100 modes are indicated.

149

$$\mathbf{w} \times \mathbf{T} = \boldsymbol{\alpha}(t)\ \mathbf{EOF} + \boldsymbol{\epsilon}^{r}(t) \tag{1}$$

150 The EOF decomposition was applied to the tropical Pacific using the learning period. As noted
151 by Kaplan et al. (1997), the variance spectrum of the EOFs could decrease too steeply due to the
152 initial coarsening of the field, leading to a strong constraint and low error estimates. To address
153 this issue, energy redistribution was performed according to equation (19) in Kaplan et al. (1997).
154 The patterns of variability are associated with the tongue in the equatorial band, extending from the
155 east to the central Pacific and the warm pool in the western Pacific (Fig. 1 a and 1c). The transect

(Fig. 1c) displayed a concentration of correlation in the subsurface layers, primarily within the first 200 m. Such patterns are typical of Eastern Pacific El Niño events (Kao and Yu 2009). The second mode captures the information on the variability of Central Pacific El Niños (Kao and Yu 2009) (Fig. 1 b and d). However, owing to the orthogonality condition of the EOFs, the physical interpretation attributed to each mode was partially flawed, potentially leading to an overlap of phenomena. Thus, the combination of these two leading EOFs is required to express the diversity of El Niño events (Trenberth and Stepaniak 2001). The following modes held lesser physical significance. The third mode was not easily interpreted (Fig. 1 e and f). The following modes explained on average 0.23% of the initial variance, (Fig. 1 g). Nonetheless, they remained valuable for consistently conveying information from observations to unsampled locations, even at depth. A fundamental assumption of our study was that these correlation distributions, represented by the eigenvectors, were robust and remained applicable in the historical context.

Second, RedAnDA differs from classical Data Assimilation schemes in that it replaces an explicit dynamical model with the Analog Backward Prediction, also known as Analog Backcast. It temporally propagates the interpolated information, offering informative *a priori* estimates of the ocean state, which are subsequently updated using available observations at corresponding times.

The use of analog prediction as a dynamical model offers several advantages, including eliminating the need to run Ocean General Circulation Models and reducing computational costs. However, it requires a dataset representative of the system's dynamics, ideally extensive enough to encompass all potential system states. This dataset, termed a catalog, includes pairs of state vectors separated by the same time offset dt, with the first element known as the analog and the second as its temporal predecessor. The catalog can be constructed from observational data or numerical simulation results. In this study, the analogs consisted of temporal coefficients $\boldsymbol{\alpha}(t)$ and their predecessors $\boldsymbol{\alpha}(t-dt)$ over the learning period, i.e., the Argo period.

A classical approach to seek analogs of a current state is by searching for its k nearest neighbors in the state space defined by a Gaussian kernel:

$$g(t,t') = \exp(-\delta^2|\alpha(t) - \alpha(t')|^2), \tag{2}$$

where t represents the time when the backcast is performed and t' is another time for which a record exists in the catalog. $|\ .\ |$ denotes the Euclidean distance, and δ is set as the inverse of

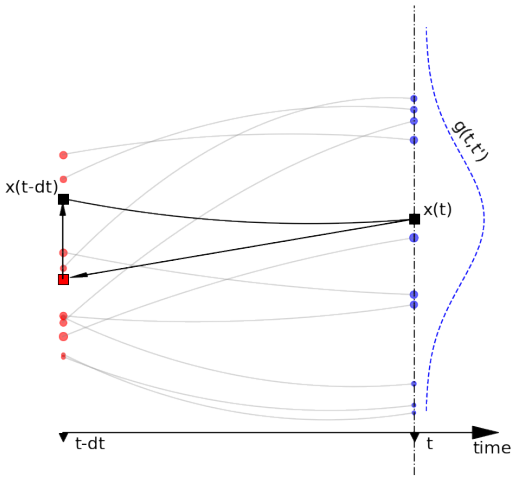


FIG. 2. The process of analog backcast prediction. The black square $x(t)$ represents the current state for which backward evolution is computed. The blue dots represent the analogs, and the red dots represent their temporal predecessors. Here, the size of the analog sample is $k = 10$. The size of each dot reflects the weighting according to similarity, as determined by the function $g(t,t')$ from equation 2, represented by the blue hatched line. The red square represents the first prediction, which results from weighted linear regression. Finally, the black square of $x(t-dt)$ represents the final result, drawn randomly around the first prediction.

the median of the Euclidean distances of the sample of the k nearest neighbors. The value of k was set to 170. This value could not be significantly increased due to limitations imposed by the catalog size n, as the quotient $k/n = 170/252$. However, this choice was expected to ease strong convergence of the linear regression in the Analog Backcast, albeit at the expense of increased computation time.

As the size of the state space, i.e., the number of EOFs, increases, the search for analogs rapidly becomes more difficult, a phenomenon known as the "Curse of Dimensionality" (**?**). This complexity makes the reduced space approach convenient in this case.

At time t, when the analog backcast $\mathcal{A}$ is applied to obtain an estimate of $\boldsymbol{\alpha}(t-dt)$, the k analogs of the current state $\boldsymbol{\alpha}(t)$ and their predecessors are selected. Each pair of analog and predecessor is then weighted according to equation (2). Assuming that both selections belong to normal distributions, a linear fit is performed, following the weighted least squares method of

Cleveland (1979). The obtained linear regression is used to extend $\boldsymbol{\alpha}(t)$ at $t - dt$, providing a first estimate of the backcast (see Fig. 2). The final backcast is drawn from the normal distribution, with the mean being the first estimate and the spread being the weighted covariance of the differences between the linearly extended analogs and their respective predecessors. This randomness reflects the uncertainties in the prediction of the analog method, which are partially controlled by the chaotic dynamics of the system. When it is observed that the linear fit efficiently links the most important analogs and their associated predecessors, the prediction is kept around its first estimate. Otherwise, the final random draw would lead to a more dispersed prediction around the first estimate. These variations in the dynamics of the system are reflected by changes in the persistence of $\boldsymbol{\alpha}$ samples. This adaptive approach is an advantage over sequential interpolation techniques that determine temporal persistence mostly on *a priori* statistics. The uncertainty of the analog prediction is given by a Monte Carlo procedure, to determine the spread of the different trajectories accessible from a set of given states, relatively close in the state space.

EOFs and temperature climatology were calculated over the Argo period, represented by the ISAS20 (In Situ Analysis System) temperature field (Gaillard et al. 2016). Associated $\boldsymbol{\alpha}(t)$ values between 2002 and 2020 were used in the catalog of RedAnDA.

The reduced set of $\boldsymbol{\alpha}(t)$ between 1950 and 2000 was then reconstructed for each month, by integrating the available *in situ* observations at time t with the Analog Backward Prediction $\mathcal{A}$, from time $t + dt$ of the analysis, where dt represents one month. The assimilated profiles came from EN4.2.2 dataset (Good et al. 2013).

In this study, sequential RedAnDA was executed in a backward temporal fashion. By starting from the state $\boldsymbol{\alpha}(t + dt)$, $\boldsymbol{\alpha}(t)$ is determined through the application of the Analog Backcast $\mathcal{A}$. A notable distinction from forward-time analysis was observed during the initialization phase, wherein the initial $\boldsymbol{\alpha}$ values were assessed based on varying quantities of available observations. This approach allowed for the transmission of comprehensive information from recent years backward in time, contrasting with the conventional forward methodology which propagates sparser historical data into the present. The RedAnDA scheme is as follows:

$$\boldsymbol{\alpha}(t) = \mathcal{A}\{\boldsymbol{\alpha}(t + dt), \epsilon^m(t + dt)\}, \tag{3}$$

$$\mathbf{y}(t) = \boldsymbol{\mathcal{H}}(t)\ \alpha(t) + \boldsymbol{\epsilon}^{o}(t). \tag{4}$$

The Analog Backcast $\mathcal{A}$ uncertainty is denoted as $\boldsymbol{\epsilon}^{m}$. The observed data are represented by $\mathbf{y}$, while $\boldsymbol{\mathcal{H}}(t)$ refers to the linear interpolation operator linking the EOF grid points to the closest observation points available at time t. The uncertainty associated with the observations is denoted as $\boldsymbol{\epsilon}^{o}(t)$.

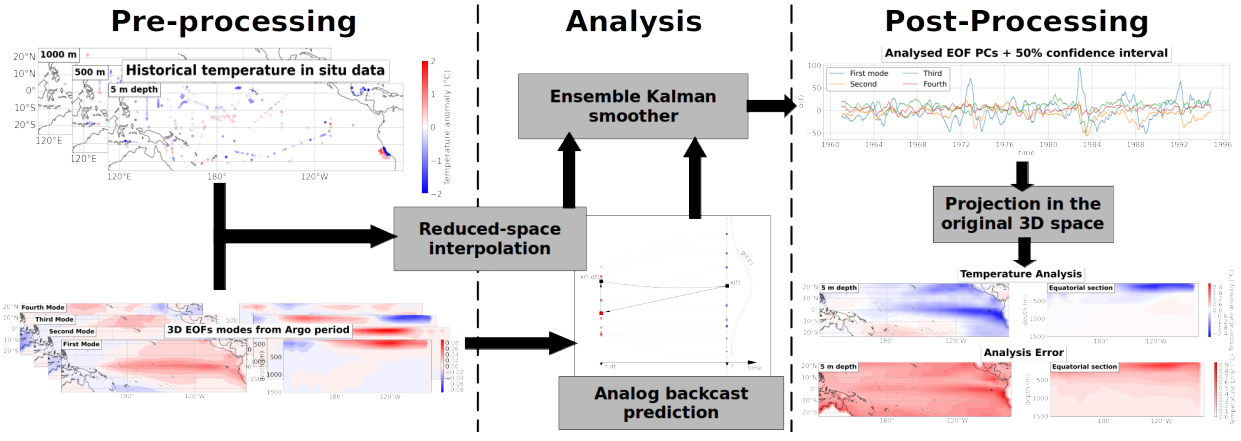


FIG. 3. RedAnDA analysis in the middle column, initiated by an initial temperature product drawn from a learning period and pre-processing of observation data shown in the left column. The post-analysis steps lead to the final gridded temperature product, in the right column.

The Reduced Space Interpolation method proposed by Kaplan et al. (1997) is consequently employed to define $\boldsymbol{\mathcal{H}}(t)$ and $\boldsymbol{\epsilon}^{o}(t)$, with detailed elaboration provided later in the section.

The analysis proceeds in accordance with the Ensemble Kalman Smoother, augmented with the Analog Backcast as depicted in Fig. 1 of **?**. Within the Ensemble variant of the Kalman filter, $\boldsymbol{\alpha}(t)$ and $\mathbf{y}(t)$ are treated as stochastic variables, following normal law distributions whose mean value and variance are determined via a Monte Carlo procedure. The variance serves to estimate the uncertainty associated with the variable.

Once $\boldsymbol{\alpha}(t)$ has been reconstructed for all desired time steps, the spatio-temporal temperature anomaly field is subsequently derived through re-projection into real space, utilizing the EOF-patterns. The whole RedAnDA procedure is summarized in figure 3. The Ensemble Kalman filter process is schematized in figure 2. The ensemble size in the Kalman Smoother was set to

500 members to ensure a more reliable calculation of the mean and covariance at each time step of the analysis. Increasing the ensemble size primarily impacted the analysis time.

Before applying RedAnDA to real observations, its efficiency is assessed through an OSSE (**?**). We apply it to the monthly output from the Oceanic Chaos – Impacts, Structure, Predictability (OCCIPUT) project (Penduff et al. 2014; Bessières et al. 2017). The simulation spans the period from 1961 to 2015. The 408 months between 1961 and 1994 are selected for reconstruction, with synthetic observations being the only source of information during this period. The 252 months over 1995-2015, for which full grid temperature maps are available, provide analogs and predecessors for the Analog prediction. The full grid maps of the reconstruction period are further used to qualify the target truth. The way these periods are split is designed to mimic the real situation where we have a recent and short Argo period with good sampling and a longer pre-Argo period, for which fewer observations are available.

In the OSSE, for comparison with the RedAnDA method, a standard Optimal Interpolation method is also implemented, following Good et al. (2013). We call it OI. It reconstructs monthly temperature maps sequentially, from 1961 to 1994, using an *a priori* monthly climatology built during the learning period, to which ~82% of the previous month's analyzed anomaly is added.

To lighten the calculations and focus on the large spatial scale, all gridded products, for the OSSE and for real reconstruction, are coarsened down to 3°x3° by horizontal averaging. Depth levels are restricted to 5, 100, 300, 500, 700, 1000, and 1500 m. The tropical Pacific is delimited here between 120°E and 70°W, 25°S and 25°N. EOFs are calculated from the resulting temperature field. For the space reduction, the number of retained EOFs was fixed at L= 10, ensuring that approximately 65% of the total variance of the initial true temperature field was explained. However, the primary EOFs primarily accounted for the high variability of the surface. To capture variability at depth, additional modes were required. With L=10, temperature signals at depths of 5, 100, 300, 500, 700, 1000, and 1500 meters were reconstructed by approximately 67%, 73%, 44%, 27%, 22%, 23%, and 20%, respectively.

In practice, $\boldsymbol{\epsilon}^o(t)$ from equation (4) was computed by combining $\boldsymbol{\epsilon}^r$, representing the uncertainty of the truncated EOFs, and an empirically derived representativeness standard deviation. This standard deviation reflected the square root of the variance of all EN4.2.2 profiles found in large horizontal boxes, at each depth. For the real observation uncertainty, $\boldsymbol{\epsilon}^o(t)$ also encompasses

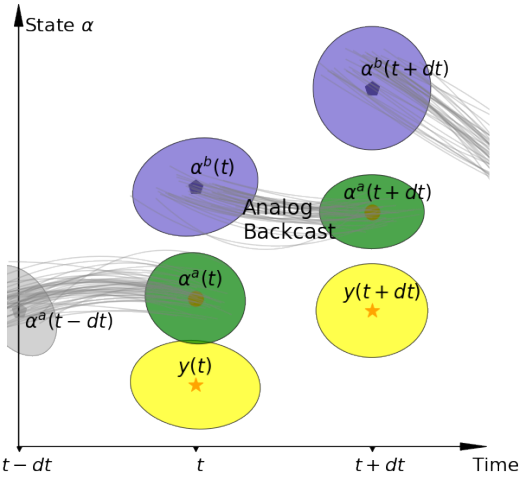


FIG. 4. Schematic diagram of the Kalman Filter process within the Data Assimilation method. Beginning with the analyzed state α^a resulting from the assimilation of observations (depicted in green), the analog backcast was executed to provide an estimate of the state at the previous time step α^b (represented in blue). Subsequently, this estimated state is updated using the *in situ* observations y (depicted in yellow), and this iterative process continues until the entire period under consideration has been processed. Each variable is depicted with its respective uncertainty, symbolized by the area of the colored shapes, which is correlated with the spread of the ensemble member trajectories (illustrated by the gray lines).

instrumental error, as assessed following **?**. This error ranged from 0.3°C for MBTs and moorings to 0.002°C for the most precise CTDs.

While the sampling of observations was generally adequate to support monthly analysis, during the early period of the time series, the availability of profiles could be limited, leading to potentially significant spurious fluctuations at large scales. To address this issue, a procedure was implemented to automatically extend the temporal window for assimilating observations. The condition for this extension was to have a minimum of 5 000 observations before resuming the analysis. This threshold was chosen because its magnitude was similar to the average number of monthly available observations in the most recent years of the analysis. At each time step, care was taken to ensure that the temporal window did not exceed the most energetic time periods of the selected EOFs, thereby preventing damping of the interpolated result.

3. Data

The ISAS20 (In Situ Analysis System) gridded product was chosen to represent the Argo period (Gaillard et al. 2016). This product provides monthly 0.5°x0.5° gridded temperature and salinity fields, constructed using an Optimal Interpolation method. It encompasses 187 standard depth levels distributed between 0 and 5500 meters and covers the period from 2002 to 2023. It constitutes an analysis of *in situ* measurements, aiming to preserve the time and space sampling capabilities of the Argo network (Kolodziejczyk et al. 2021). Therefore, two correlations scale were used: 300 km and 4 times the deformation radius (Gaillard et al. 2016).

The assimilated profiles for reconstructing past periods came from the EN4.2.2 dataset (Good et al. 2013). This dataset represents the latest version of the Met Office Hadley Centre's 'EN' dataset series, offering quality-controlled potential temperature profiles from 1900 to the present day. Prior to the Argo period, most profiles were sourced from the World Ocean Database 2018 (WOD18) (Boyer et al. 2006), and from the Global Temperature and Salinity Profile Program (GTSPP) (Sun et al. 2010). Notably, no Argo profiles were included in the latter dataset. In the synthetic profiles of the OSSE as well as in EN4.2.2, only depths where the temperature observation had a quality flag of 1 are retained. Additionally, since these depths varied from profile to profile, a vertical linear interpolation was conducted to obtain the profiles at the target reconstruction grid levels. No extrapolation was performed. To filter out high frequencies in observations from fixed tropical moorings (TAO/TRITON array in the Pacific), it was decided to average the time series over 15-day periods for each identified buoy. Super-profiles were then considered at the center of the time slice, reducing the number of initial profiles from 832 703 to 648 786. In figure 5, the number of available observations, between 1950 and 2000, appears to increase over time, with a substantial improvement in sampling in the nineties. Two minima are found in the fifties and eighties, suggesting that reconstruction during these periods may have been less constrained.

In the OSSE framework, the Oceanic Chaos – Impacts, Structure, Predictability (OCCIPUT) project was utilized (Penduff et al. 2014; Bessières et al. 2017). The OCCIPUT project offers an ensemble of 50 realizations of a global 1/4° ocean/sea ice simulation, based on the NEMO 3.5 model (Penduff et al. 2014; Bessières et al. 2017), alongside synthetic profiles extracted at specific times and locations corresponding to the historical database. For this study, only one member of

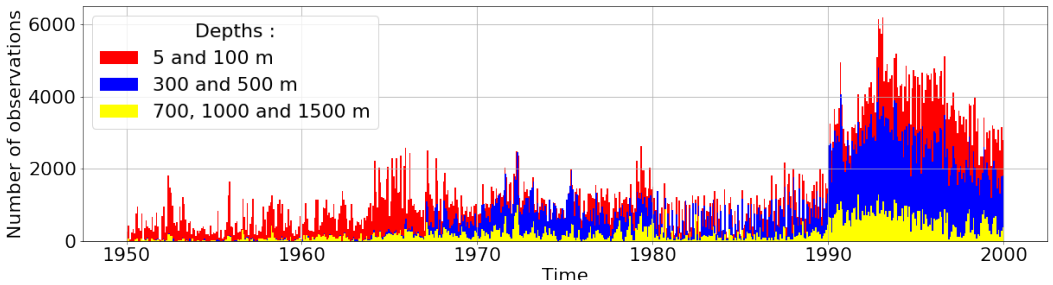


FIG. 5. Monthly histogram of available EN4.2.2 temperature observations in the tropical Pacific, depending on the depth, between 1950 and 2000. In red, observations at 5 and 100 meters depth are counted. In blue, observations at 300 and 500 meters depth are counted. The number of observations at 700, 1000 and 1500 meters depth is represented in yellow.

the model output and synthetic profiles were used. The simulation spans the period from 1961 to 2015, employing the DFS atmospheric forcing over the same period (Dussin et al. 2016).

To compare RedAnDA's temperature reconstruction with state-of-the-art reanalysis products since the 1950s, five products were introduced. Three of these are optimal interpolation products, while the other two are derived from data assimilation with numerical models:

- ISHII: Version 6.4.0 of a historical objective analysis of temperature data provided by the Japan Marine Science and Technology Center. ISHII utilizes *in situ* observations from the World Ocean Database 2005 (WOD05), the Global Temperature and Salinity Profile Program (GTSPP), Centennial *in situ* Observation Based Estimates (COBE), and ARGO buoys (**?**). It compiles monthly 1°x1° temperature fields on 24 depth levels.

- EN4: A gridded product from the Hadley Centre of the UK Meteorological Office. EN4 employs an optimal interpolation technique, with *in situ* profiles from the EN4.2.2 dataset analyzed using a smoothed 1971–2000 field from the EN2 version (Good et al. 2013). It provides monthly temperature fields from 1900 to the present day over 42 depth levels.

- IAP product: Provided by Ensemble Optimal Interpolation of the CODC-GOSD dataset, it utilizes CMIP5 models to reconstruct global and gridded temperature at a monthly period from 1940 to the present day, at a spatial resolution of 1°x1° in the upper 2000 m, over 41 levels (**??**) .

- ESTOC dataset: Developed by the Japan Agency for Marine-Earth Science and Technology (JAMSTEC) and Kyoto University, ESTOC utilizes a four-dimensional variational data assimilation method with the GFDL Modular Ocean Model (MOM3). It assimilates quality-controlled EN3 *in situ* profiles and sea-surface dynamic-height data from Aviso altimetry products (**??**). The 1°x1° monthly result covers the period from 1956 to 2009 over 46 vertical levels.

- ORAS4: Provided by the Copernicus Climate Change Service, ORAS4 combines model data with observations using the NEMO model and the 3D-Var assimilation system called NEMOVAR (**?**). It assimilates *in situ* profiles from the EN3 dataset, sea-surface temperature from the OIv2 product, and altimeter sea level anomalies from the AVISO dataset. ORAS4 covers the period from 1959 to the present day on 42 depth levels and with a horizontal spatial resolution of 1°x1°.

4. Validation using numerical simulations and OSSE

The tropical Pacific temperature fields were reconstructed within the OSSE framework, with the assimilation of synthetic observations available between 1961 and 1994. The 1995-2015 OCCIPUT model (Penduff et al. 2014; Bessières et al. 2017) fields served as a learning period for the Analog backcast.

Our validation initially focused on selected depths (around 100 m, 500 m, and 1000 m), analyzing the temporal evolution of temperature anomalies averaged over specific sub-regions chosen to highlight the key features of both RedAnDA and OI products, as well as the model truth.

We first present the reconstructions in the tropical Pacific at 100-m depth level (Fig. 6a, b, and c). During the 1973 El Niño, as identified in the Niño 3.4 time series (Fig. 6d), the temperature field appeared consistently reconstructed with RedAnDA (Fig. 6a and c). However, the OI reconstruction lacked the necessary observations to successfully capture the spatial distribution of the warm and cold anomalous patterns of the El Niño event (Fig. 6b). Despite persistence from the initial guess, which allowed the emergence of a warm signal around 160°E, the region remained overall too cold in the absence of surface observations. In contrast, the cold anomaly in the North Equatorial Countercurrent around 5-10°N was satisfactorily reconstructed with RedAnDA (Fig. 6a and c).

Due to an insufficient number of observations and inaccurate *a priori* correlation scales, OI did not reproduce such fine structure (Fig. 6b and c).

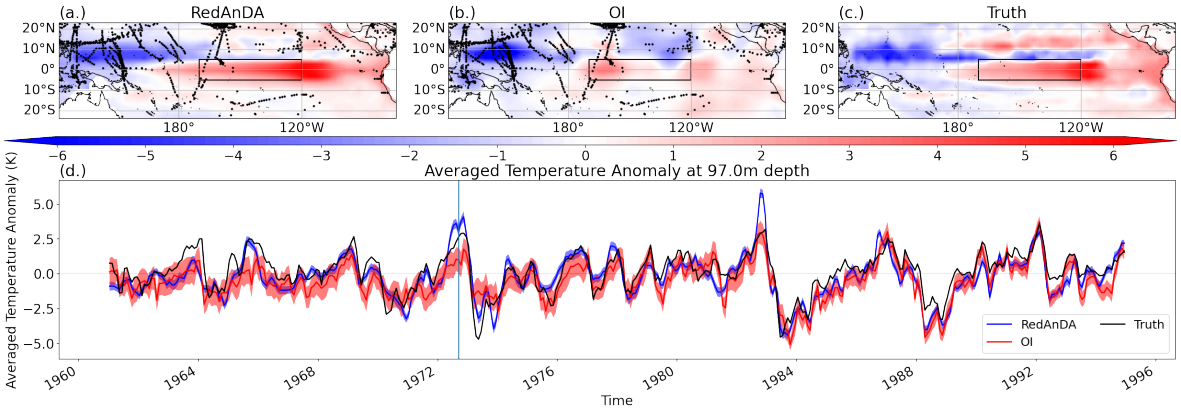


FIG. 6. (a, b, c) Temperature anomalies at around 100-m depth for RedAnDA, OI, and OSSE truth in September 1972. The black dots indicate synthetic observations available at the given depth and time. The black box in panels a, b, and c shows the region where the temperature is calculated to obtain a pseudo-Niño 3.4 index. (d) Temporal variations in this average for RedAnDA, OI, and OSSE truth. RedAnDA and EN4 provide 50% confidence intervals. The blue vertical line indicates the selected month for the top snapshots.

The time series of the averaged temperature anomaly in the Niño 3.4 box, a region where an SST anomaly index is typically calculated, indicated that both methods performed better in the recent past, from 1974 to 1994, when more observations were available (Fig.5 and 6d). In this selected region, the temperature anomaly was strongly positive during El Niño events, and most of them were successfully identified in the years: 1963, 1965, 1968, 1969, 1972, 1976, 1977, 1979, 1983, 1987, 1990, 1992, 1993, and 1994. Interestingly, in the early period, the El Niño events of 1963, 1965, 1972, and 1976 were systematically underestimated by about 1°C by the OI reconstruction. Similarly, most La Niña events were identified in both products (1962, 1964, 1971, 1973, 1974, 1975, 1984, and 1988), but tended to be overestimated by 1-2°C, as in 1971, 1974, and 1988. Occasionally, such biases could have led to an overestimation of temperature, as in 1972 and 1983.

At 500-m depth, the true signal exhibited low-frequency variability that was scarcely detected by both OI and RedAnDA (Fig. 7a, b and c). This was particularly evident outside the equatorial zone, where a region with strong cold anomalies of approximately −1.2°C extended along latitudes 20°N and 20°S. From the Coral Sea around 18°S and 158°E to the eastern boundary of the Pacific, both reconstructions failed to capture the extent of the cold anomaly (Fig. 7a, b, and c). Although

RedAnDA and OI reconstructed the eastern warm and northwestern cold regions, they did so with an intensity underestimated by approximately 0.2°C. Additionally, in the eastern tropics, the La Niña pattern at 500-m depth in RedAnDA was more realistic compared to that in OI. For instance, RedAnDA was able to identify the equatorial colder temperatures near the South American coast.

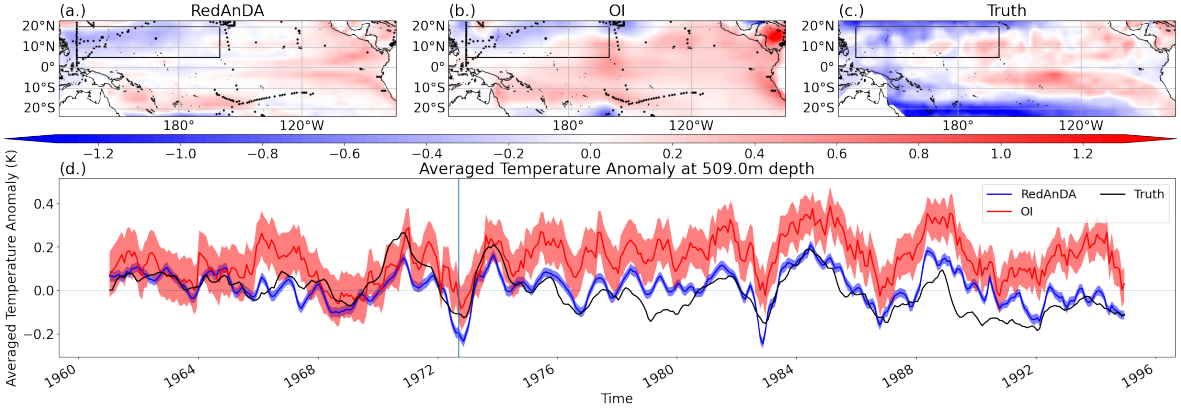


FIG. 7. (a, b, c) Temperature anomalies around 500 m depth for RedAnDA, OI, and OSSE truth in September 1972. Black dots indicate synthetic observations available at the given depth and time. The black box in panels a, b, and c, now located between 5-20°N, 130°E-160°W, shows the region where the temperature is calculated to estimate the signature of the westward North Equatorial Subsurface Current (NESC) (**?**). (d) Temporal variations in this average for RedAnDA, OI, and OSSE truth. RedAnDA and OI provide 50% confidence intervals. The blue vertical line indicates the selected month for the top snapshots.

The temporal variations illustrated RedAnDA's capability to reconstruct the averaged variability in the selected box in the northwest of the domain, between 5°-20°N and 130°E-160°W (Fig. 7c and d). The choice of the box was made to capture the highest regional variability at 500-m depth, which differed from the area at 5 m depth. This variability is associated with the North Equatorial Subsurface Current (NESC), which varies interannually along with ENSO. It flows stronger during La Niña events and weaker during El Niño (**?**). However, some variations appeared excessively pronounced in the RedAnDA product, such as in 1967 with a difference of 0.1°C, or in recent years, while others were underrepresented with similar magnitude, such as in 1971, around 1980, and between 1989 and 1992 (Fig. 7d). In contrast, OI adequately represented the period 1968 to 1974 and exhibited variations that correlated with RedAnDA most of the time. However, it suffered from a drift that led to an overestimation of temperature by approximately 0.2°C in the

selected region over time. Moreover, the OI signal was notably noisy, due to the irregular sampling of observations.

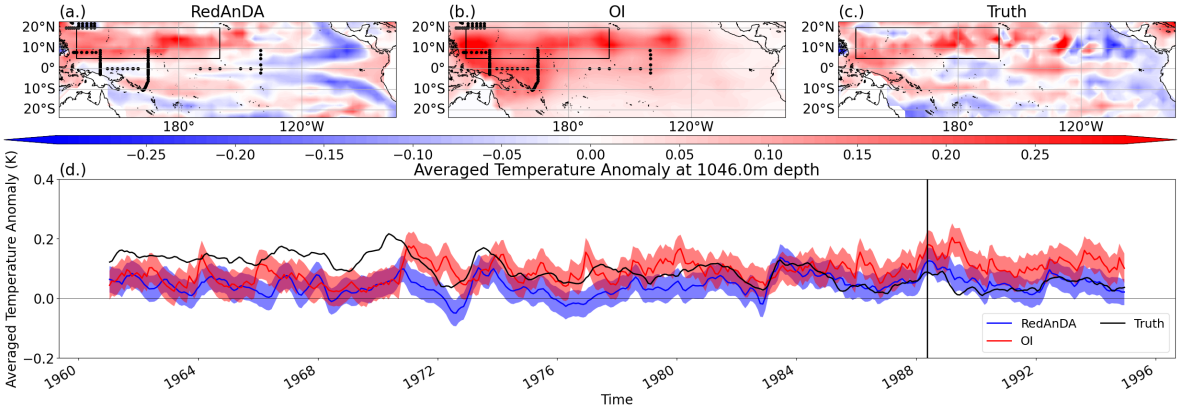


FIG. 8. (a, b, c) Temperature anomalies around 1000 m depth for RedAnDA, OI, and OSSE truth in May 1988. Black dots indicate synthetic observations available at the given depth and time. The black box in panels a, b, and c, between 5-20°N, 130°E-160°W, shows the region where the temperature is averaged to assess the variability associated with the North equatorial deep jets (**???**). (d) Temporal variations in this average for RedAnDA, OI, and OSSE truth. RedAnDA and OI provide 50% confidence intervals. The blue vertical line indicates the selected month for the top snapshots.

Around 1000 m depth, the number of observations diminished significantly (Fig. 5 and 8). It impacted the OI reconstruction, particularly east of 120°W (Fig. 8b). In the northwest tropical Pacific, the OI captured the warm zonal pattern along 15°N with an anomaly of 0.25°C (Fig. 8b and c). In contrast, RedAnDA extrapolated more signal from analogs, aided by three-dimensional EOFs connecting the surface to the depths (Fig. 8a). Although its reconstruction may have been insufficient to represent small-scale structures. From 120°E to 120°W in the northern tropics and between 150°E to 100°W in the south, RedAnDA successfully retrieved the warm extensions, along with cold regions of approximately −0.2°C east of 120°W along the coast. Along the equator, the most significant differences between RedAnDA and Truth were observed, with dissimilar warm and cold patterns (Fig. 8a and c).

Considering the time series in the selected box, OI had difficulties capturing the most significant temperature variance (Fig. 8d), exhibiting high-frequency oscillations not coherent with the time series derived from the Truth. This is likely attributed to sparse and uneven distribution in data sampling. Initially, in 1961, OI closely aligned with RedAnDA temperature estimates, but then

diverged from model Truth and RedAnDA by up to 0.1°C, as persistence introduced a drift. Its correction was impeded by either an insufficient number of observations or an overemphasis on background signals.

RedAnDA's temperature time series correlated more closely with the truth, especially after 1983. Before 1983, the oscillations were recovered, but often too weakly and biased, as the true signal presented a drift. Before 1970, the difference due to the drift was greatest, and RedAnDA's reconstructed oscillations correlated less closely with the truth, with a difference of 0.05°C in 1961.

To evaluate the quality of the uncertainty estimation, we computed the probability coverage. It is evaluated as the proportion of the true signal within a given confidence interval of the reconstruction. The confidence interval was determined by the normal quantile q, set to approximately 0.67 for a 50% confidence interval. This metric assessed the reliability of uncertainty estimates in capturing differences between the true and reconstructed signals. A proportion close to 50% indicated proper uncertainty estimation, while deviations from 50% indicated potential underestimation or overestimation of uncertainty. For more details about this metric, see **?**.

$$\text{Probability Coverage} = P\left[\mathbf{T}(t) - q \times \boldsymbol{\sigma} \geq \mathbf{T}^{\text{true}}(t) \geq \mathbf{T}(t) + q \times \boldsymbol{\sigma}\right], \tag{5}$$

Generally, both methods captured less than the desired amount of truth within their confidence intervals (Fig. 9e, f, g and h). For OI, the uncertainty was basin-wide underestimated. Spatially, only 30% of the truth was captured within the confidence interval over most of the domain where *a priori* variances were the highest (Fig. 9c, d, g, and h). Below 300 m, estimates were the least reliable, which could be attributed to observed drifts (Figs. 7d and 8d). At depth, in the southern half of the basin, the reconstruction of anomalies was better, leading to higher inclusion percentages, between 30-40% (Fig. 9g).

The RedAnDA uncertainty was linked to the space reduction, the sampling of observations, and the dispersion of members during the backcast prediction. The confidence interval was small, given the variability of the signal under consideration and the EN4 uncertainty (Figs. 6d and 7d). Before 1970, the confidence interval was wider, but still failed to capture the true signal when

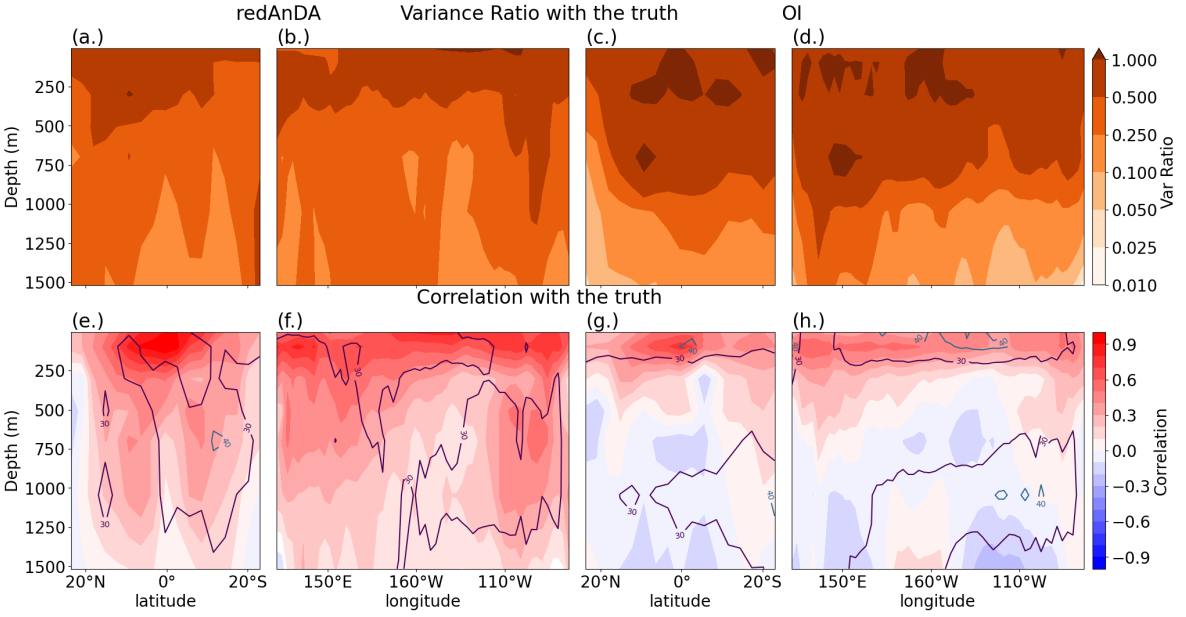


FIG. 9. Meridional and zonal mean sections depicting the ratio of reconstructed variance to true variance (first row) and the correlation of the two analysis products with the OSSE truth (second row). In the panels of the second row, purple and blue contours indicate the 30% and 40% values in probability coverage. For the variance ratio, note the logarithmic color scale. The two left columns represent the RedAnDA products, while the two right columns represent the OI products.

the drift was significant (Fig. 8d). Overall, the results were overconfident at the surface, as the inclusion of the truth within the confidence interval was less than 30 %. The quantification of uncertainty was better at depth (Fig. 9e and f).

Supplementary performance scores were calculated using several metrics. These metrics included the variance ratio and correlation defined as follows:

$$\text{Variance Ratio} = \frac{\overline{\mathbf{T}(t)^2}}{\overline{\mathbf{T}^{\text{true}}(t)^2}}, \tag{6}$$

The variance ratio compared the variance of the reconstructed signal $\mathbf{T}$, to the variance of the true signal $\mathbf{T}^{\text{true}}$. A variance ratio of 1 indicated that the reconstructed variability matched the true variability in energy.

$$\text{Correlation} = \frac{\overline{(\mathbf{T}(t) - \mathbf{T}^{\text{true}}(t))^2}}{\sqrt{\overline{\mathbf{T}(t)^2} \times \overline{\mathbf{T}^{\text{true}}(t)^2}}}, \tag{7}$$

The correlation measured the coherence between the true and reconstructed signals. It ranged from −1 and 1, where −1 indicated opposite phase variations, 1 indicated similar variations, and 0 indicated no consistency between the signals.

Both analyses demonstrated the capability to inject a satisfactory level of variability into the upper 300 m, where they also exhibited satisfactory correlations with the model Truth (Fig. 9a-h). In the warm pool to the west along the equator and in the warm extensions of El Niño, between 160°W and the eastern basin, the effectiveness of the RedAnDA reconstruction was notably evident (Fig. 9a, b, e, and f). This outcome was anticipated, given that the principal modes of EOFs were linked to the variability of these regions. In contrast, OI exhibited satisfactory variance ratios and correlations for reconstruction, primarily concentrated in regions with better sampling (Fig. 9c, d, g, and h).

At depths exceeding 300 meters, OI continued to inject information but displayed weak or negative correlations (ranging from 0 to −0.2) with the ground truth (Fig. 9c, d, g, and h). The signal reconstructed by RedAnDA exhibited more consistency with the truth, displaying positive correlation reaching up to 0.5 (Fig. 9a, b, e, and f). However, at the northern and southern boundaries, below 300 m, the reconstruction's accuracy was less evident, owing to drifts in the model truth that neither RedAnDA nor OI identified (Figs. 7d and 8d).

5. Results

The regional reconstruction of temperature was conducted using real observations spanning from 1950 to 2000 in the tropical Pacific. In this section, the performance of RedAnDA is compared with other temperature fields derived from various analyses.

Firstly, Sea Surface Temperature (SST) reconstruction maps were examined for the intense 1982 El Niño event (Fig. 10a-f), during which observation availability was not homogeneous. All reconstructions depicted the warm tongue extending 2°C to the east and the cold region to the west. Along 10°N, between 180°E and 150°W, the cold anomaly was detected by all methods.

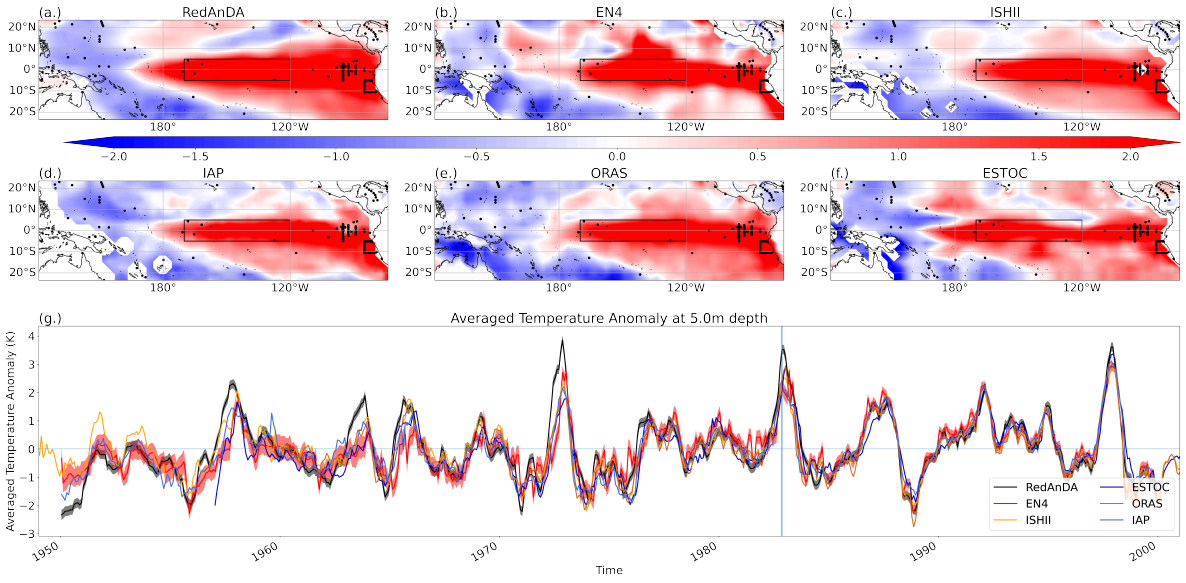


FIG. 10. Temperature anomalies at 5-m depth in November 1982 for various datasets, including RedAnDA, EN4, ISHII, IAP, ORAS, and ESTOC. The upper two rows display snapshots of temperature anomalies for each dataset, with black dots indicating available observations at the given depth and time. The observation positions represented here may not precisely match those assimilated by analyses other than RedAnDA. The black box highlights the region where temperature was averaged to obtain a Niño 3.4 index. In the lower panel, the temporal variations of this average temperature anomaly are depicted. Both RedAnDA and EN4 provided 50% confidence intervals. The blue vertical line denotes the time selected for the upper snapshots. To enhance visibility, the RedAnDA curve is depicted in black.

However, around 120°W-15°N in the northern Tropical Pacific, there was disagreement among the reconstructions regarding the recovery of the cold anomalies (Fig. 10 a).

The Niño 3.4 index exhibited good agreement among all time series (Fig. 10 g). ENSO events in RedAnDA appeared to be slightly stronger (up to 1°C) compared to other products, although previous validation results in section 4 demonstrated its ability to accurately recover El Niño events.

In the 1950s, intense cooling was observed in RedAnDA, which was not observed in the other analyses. Error estimates for both EN4 and RedAnDA decreased over time as the number of observations increased. The error interval for RedAnDA was notably thinner compared to EN4, potentially indicating an underestimated error in the reconstruction.

At a depth of 300 meters, in February 1978, the various methods yielded less consistent results (Fig. 11a-f). In the southwest of the domain around 180°E, there was a consensus regarding the presence of a cold anomaly pattern, approximately −2°C, with, on its southern and northern flanks, warm anomalies ranging between 0.5 and 1°C. However, in the eastern Pacific, significant disparities emerged despite the presence of some observations near the coast. Around 120°W, between 15°S and 10°N, IAP, ESTOC, and RedAnDA appeared to agree on a warm anomaly pattern whereas ORAS tended toward a colder anomaly (Fig. 11a, d, e, and f). In this region, RedAnDA diverged from EN4 and ISHII, but these reconstructions, lacking sufficient observational data, exhibited low energy (Fig. 11b and c).

In the box (Fig. 11a-f), located in the southwest of the study area, between 150°E - 160°W and 5 - 15 °S, temporal fluctuations relate in part to the Pacific South Equatorial Current (SEC). This current undergoes interannual fluctuations in its intensity related to ENSO, as it flows stronger a few months following an El Niño event and weaker after a La Niña event (**?**). Concerning the estimation of this variability, after the 1970s, the different time series are aligned (Fig. 11 g). In 1979, RedAnDA reconstructed a significant cold event with a magnitude around 1°C, which was less pronounced in other methods, estimating its magnitude at −0.5°C. This discrepancy arose from the cold extension between 5°-15°S and 150°E to 120°W and the warm southern pattern centered at 15°S and 165°E, which were larger and narrower, respectively, compared to other methods (Fig 11a-f).

A similar discrepancy occurred in 1959. In the 50s, RedAnDA assimilated more observations over a wider time window (see section 2), resulting in a smoother analysis during this period than in other analyses. Except for the 1959 event, the variations correlated with the highest peaks in IAP. However, these were likely associated with spurious assimilation of sparse observations.

The average correlation of RedAnDA with the other products indicated overall agreement (Tab. 1). ISHII and ORAS showed the closest similarity to RedAnDA, with both correlations at 0.61, while ESTOC showed the least similarity with a value of 0.47. These close scores demonstrated that RedAnDA aligned with the state-of-the-art reconstructions. However, the existence of dissimilarity suggested that RedAnDA still provided original results, not yet obtained with any of the alternative methods. Regarding the results from the OSSE, the original reconstruction of RedAnDA in the

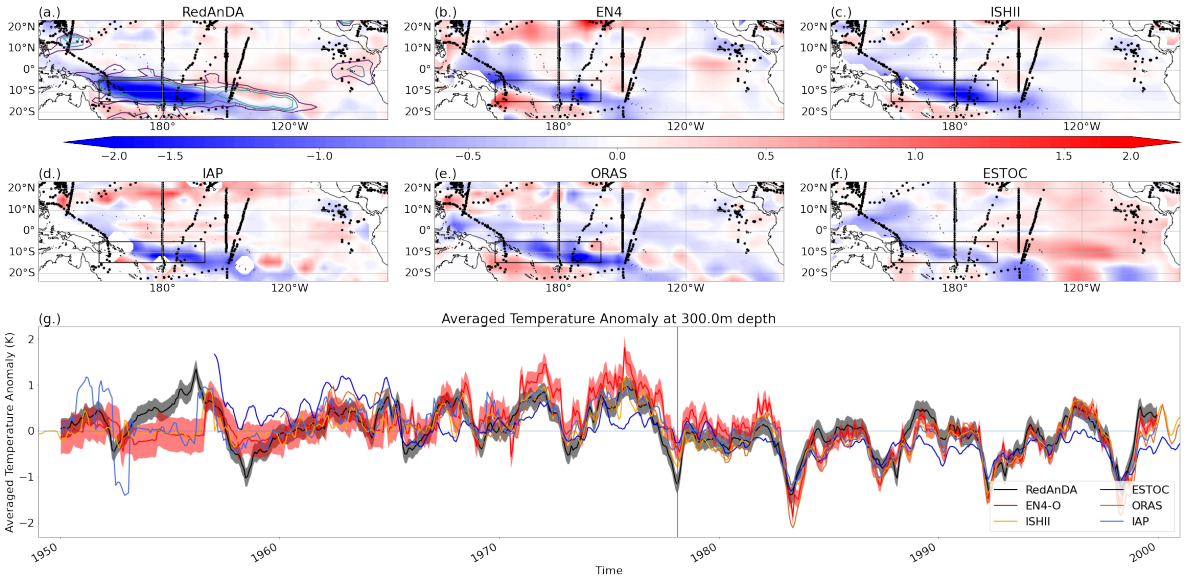


FIG. 11. Temperature anomalies at 300m depth in February 1978, including RedAnDA, EN4, ISHII, IAP, ORAS, and ESTOC. The upper two rows display snapshots of temperature anomalies for each dataset, with black dots indicating available observations at the given depth and time. The observation positions represented here may not precisely match those assimilated by analyses other than RedAnDA. The contours in (a) delimit the 0.35, 0.45 and 0.55°C temperature standard deviations to highlight the regions of high variability. The black box, now between 150°E - 160°W and 5 - 15 °S, has changed to highlight the southwestern region where temperature was averaged to visualize variability partly associated with the Pacific South Equatorial Current (SEC) (**???**). In the lower panel, the temporal variations of this average temperature anomaly are depicted. Both RedAnDA and EN4 provide their 50% confidence intervals. To enhance visibility, only these are shown. The blue vertical line denotes the time selected for the upper snapshots. To enhance visibility, the RedAnDA curve is depicted in black.

upper 500 m depth may be worth consideration as it may have reconstructed components of the variability that were undetected from other datasets.

6. Discussion & Conclusion

In conclusion, the RedAndA method represents a pioneering approach by integrating Analog backward prediction Data Assimilation with reduced-space analysis. It demonstrated successful reconstruction of the monthly three-dimensional temperature field from sparse and randomly distributed *in situ* profiles in the Pacific Ocean. In the OSSE, RedAnDA outperformed Optimal

	Correlation
ISHII	0.61
EN4	0.52
IAP	0.57
ORAS	0.61
ESTOC	0.47

TABLE 1. Comparative analysis of the RedAnDA reconstruction with reconstructions from ISHII, EN4, IAP, ORAS, and ESTOC, listed in rows from first to fifth, respectively. The scores represent the Correlation. Each comparison is conducted over the period common to both reconstructions.

Interpolation (OI) in terms of truth retrieval. Moreover, it exhibited agreement with alternative reconstructions such as ISHII, IAP, ORAS4, and ESTOC, while inferring variability where other methods lacked information and avoiding spurious results.

RedAnDA's advancements relied on its ability to extrapolate to regions with poor sampling, particularly at depth, facilitated by 3D EOFs that correlated the different vertical levels. Additionally, the analog backcast provides statistical temporal predictions that enhance the estimation of the *a priori* statistics every time step in the analysis.

However, evaluating the uncertainty associated with RedAnDA remained challenging, encompassing various sources such as the quality and sampling of observations, the predictability of the dynamics, and the unrepresented scales. Further investigation into each of these sources of error will be studied in future works.

It is worth noting that assumptions regarding the validity of climatology and EOFs during the reconstruction period, as they were calculated during the learning period may have had limitations. As demonstrated in the OSSE, these limitations could have induced some errors, notably concerning long-term changes in the climate system. Thus, estimating the associated uncertainty and understanding the relationship between temporal distance from the learning period and obsolescence are crucial areas for future research.

In the OSSE, the non-stationarity of the mean state was found to be significant, particularly at depth. Attempts to account for the changing climatology yielded unsatisfactory results. The observations were insufficient in reflecting the long-term evolution adequately, likely due to uneven sampling.

Space reduction constrained signal extrapolation at depth, where a greater number of EOFs were required to fully capture variability. As previously stated, the curse of dimensionality imposed limitations on the number of EOFs that could be utilized (**?**). Efforts were needed to enhance the EOFs to encompass more of the variability present in the initial signal. Focusing solely on a specific depth, it is feasible to adjust the relative significance of each level in the construction of EOFs through weighting, allowing for the incorporation of the most crucial modes of variability associated with the target depth.

The length of the catalog was constrained by the availability of Argo data, which might have been insufficient to represent the full spectrum of ocean dynamics. Consequently, there was a risk of underestimating the uncertainty in the analysis due to the potential redundancy of the selected analogs and the inherent replication of associated trajectories. While using a catalog composed of simulated data could have expanded its size, the primary objective was to generate an analysis solely based on real observations. This approach aimed to avoid potential bias introduced by numerical models, which could impact data assimilation despite observational constraints, particularly in regions and periods with sparse sampling (**??**). Given that the influence of such factors was not assessed in our case, it was decided not to incorporate simulated data into our reconstruction of reality.

Within the OSSE framework, it was envisaged exploring a broader catalog by incorporating additional members from the OCCIPUT ensemble (Penduff et al. 2014). However, since all 50 members were simulated by the same forced model with the same atmospheric forcing. In the case of a forced OGCM, the variability associated with ENSO reduces to forced variability. Thus, it was determined that the tropical Pacific basin, with its predominant wind forcing, might not provide a significant enhancement in dynamics for the catalog (**?**). In regions where ocean internal variability holds greater sway (e.g., the North Atlantic or the Southern Ocean), using multiple members may be crucial for reliably assessing uncertainty.

For a real observational application, the catalog, climatology, and EOFs were derived from ISAS, which already constituted an analysis of the Argo period. While the effectiveness of its optimal interpolation had been demonstrated, its limitations – such as the lack of signal in shallow coastal areas – undoubtedly constrained RedAnDA. In the OSSE, the learning period was optimal. One

method to evaluate the impact of the ISAS mapping technique would involve applying it to the simulation prior to reconstructing the past and observing the resulting modifications.

The current analysis was intended for investigations into regional temperature interannual variability and at large time scales as well. Oceanic structures that were insufficiently observed during the 20th century, notably at depth, may be characterized through the use of the new temperature product provided by RedAnDA.

To reconstruct temperature and salinity with RedAnDA in a co-varying approach, could provide new datasets of the evolution of regional water masses over the second half of the 20^{th} century. To consider their co-variability would be beneficial to salinity from the greater number of temperature observations.

An illustrative application involves utilizing RedAnDA's Pacific temperature time series within a new catalog to predict El Niño using analog forecasting. With the Precursors method (e.g., **????**), the analogs could provide new means of identifying conditions conductive to ENSO events.

Acknowledgments. Erwan Oulhen was funded by the Région Bretagne and the Université de Bretagne Occidentale. This study was led in the framework of the LEFE GMMC OASIS project (INSU/GMMC). ISAS temperature and salinity monthly gridded field products are made freely available by SNO Argo France at LOPS Laboratory (supported by UBO/CNRS/Ifremer/IRD) and IUEM Observatory (OSU IUEM/CNRS/INSU) at doi: https://doi.org/10.17882/52367. The authors express gratitude to Thierry Penduff and Jean-Marc Molines for providing OCCIPUT outputs and their invaluable assistance. Appreciation is also extended to Gaël Alory, Laurent Bertino, Sally Close, William Llovel, and Jérôme Gourrion for their constructive discussions.

Data availability statement. ISAS20 data are available through `10.17882/52367`; EN4.2.2 profiles and objective analyses are available through `https://www.metoffice.gov.uk/hadobs/en4/download-en4-2-2.html`; OCCIPUT data are available upon demand; ORAS4 are available through `https://icdc.cen.uni-hamburg.de/thredds/dodsC/oras4_temp_all.html`; IAP data are available through `http://www.ocean.iap.ac.cn/ftp/cheng/CZ16_v3_IAP_Temperature_gridded_1month_netcdf/`; ISHII data are available through `https://rda.ucar.edu/datasets/ds285.3/dataaccess/`; ESTOC are available through `https://www.godac.jamstec.go.jp/jagdas/catalog/estoc/02c/catalog.html`. RedAnDA results and codes will be provided upon demand.

References

Baldwin, M. P., D. B. Stephenson, and I. T. Jolliffe, 2009: Spatial weighting and iterative projection methods for eofs. *Journal of Climate*, **22 (2)**, 234 – 243, https://doi.org/10.1175/2008JCLI2147.1.

Bessières, L., and Coauthors, 2017: Development of a probabilistic ocean modelling system based on nemo 3.5: application at eddying resolution. *Geoscientific Model Development*, **10 (3)**, 1091–1106, https://doi.org/10.5194/gmd-10-1091-2017.

Boyer, T. P., and Coauthors, 2006: Noaa atlas nesdis 60. *World ocean database 2005*.

Bretherton, F. P., R. E. Davis, and C. Fandry, 1976: A technique for objective analysis and design of oceanographic experiments applied to mode-73. *Deep Sea Research and Oceanographic Abstracts*, **23 (7)**, 559–582, https://doi.org/https://doi.org/10.1016/0011-7471(76)90001-2.

Cleveland, W. S., 1979: Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74 (368)**, 829–836, https://doi.org/10.1080/01621459.1979.10481038, https://www.tandfonline.com/doi/pdf/10.1080/01621459.1979.10481038.

Dussin, R., B. Barnier, L. Brodeau, and J. M. Molines, 2016: Drakkar forcing set dfs5. *MyOcean Report*.

Emery, W. J., and R. E. Thomson, 2001: Chapter 4 - the spatial analyses of data fields. 305–370, https://doi.org/https://doi.org/10.1016/B978-044450756-3/50005-8.

Gaillard, F., T. Reynaud, V. Thierry, N. Kolodziejczyk, and K. von Schuckmann, 2016: In situ–based reanalysis of the global ocean temperature and salinity with isas: Variability of the heat content and steric height. *Journal of Climate*, **29 (4)**, 1305 – 1323, https://doi.org/10.1175/JCLI-D-15-0028.1.

Good, S. A., M. J. Martin, and N. A. Rayner, 2013: En4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. *Journal of Geophysical Research: Oceans*, **118 (12)**, 6704–6716, https://doi.org/https://doi.org/10.1002/2013JC009067, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2013JC009067.

Hansen, J., M. Sato, P. Kharecha, and K. von Schuckmann, 2011: Earth's energy imbalance and implications. *Atmos. Chem. Phys.*, **11**, 13 421–13 449, https://doi.org/10.5194/acp-11-13421-2011.

Kao, H.-Y., and J.-Y. Yu, 2009: Contrasting eastern-pacific and central-pacific types of enso. *Journal of Climate*, **22 (3)**, 615 – 632, https://doi.org/10.1175/2008JCLI2309.1.

Kaplan, A., Y. Kushnir, M. Cane, and M. Blumenthal, 1997: Reduced space optimal analysis for historical data sets: 136 years of atlantic sea surface temperatures. *J. Geophys. Res.*, **102860**, 835–27, https://doi.org/10.1029/97JC01734.

Kolodziejczyk, N., A. Prigent-Mazella, and F. Gaillard, 2021: Isas temperature and salinity gridded fields. https://doi.org/https://doi.org/10.17882/52367.

Lguensat, R., P. Tandeo, P. Ailliot, M. PULIDO, and R. Fablet, 2017: The analog data assimilation. *Monthly Weather Review*, **145 (10)**, 4093 – 4107, https://doi.org/10.1175/MWR-D-16-0441.1.

Penduff, T., B. Barnier, L. Terray, L. Bessières, and G. Sérazin, 2014: Ensembles of eddying ocean simulations for climate.

Shea, D., 2013: National center for atmospheric research staff (eds). The Climate Data Guide: Empirical Orthogonal Function (EOF) Analysis and Rotated EOF Analysis. Retrieved from urlhttps://climatedataguide.ucar.edu/climate-data-tools-and-analysis/empirical-orthogonal-function-eof-analysis-and-rotated-eof-analysis.

Sun, C., and Coauthors, 2010: The data management system for the global temperature and salinity profile programme. *Proceedings of OceanObs*, **9**, 86.

Trenberth, K. E., and D. P. Stepaniak, 2001: Indices of el niño evolution. *Journal of Climate*, **14 (8)**, 1697 – 1701, https://doi.org/10.1175/1520-0442(2001)014<1697:LIOENO>2.0.CO;2.

Zhen, Y., P. Tandeo, S. Leroux, S. Metref, T. Penduff, and J. L. Sommer, 2020: An adaptive optimal interpolation based on analog forecasting: Application to ssh in the gulf of mexico. *Journal of Atmospheric and Oceanic Technology*, **37 (9)**, 1697 – 1711, https://doi.org/10.1175/JTECH-D-20-0001.1.