

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 AND FLORIAN SÉVELLEC^{a,c}

^a Univ. Brest, CNRS, IRD, Ifremer Laboratoire d'Océanographie Physique et Spatiale (UMR 6523), IUEM, Brest, France

^b IMT Atlantique, Lab-STICC, UMR CNRS 6285, Brest, France

^c Odyssey, Inria/IMT/CNRS, Rennes, France

^d RIKEN Center for Computational Science, Kobe, Japan

(Manuscript received 14 November 2023, in final form 30 August 2024, accepted 15 October 2024)

ABSTRACT: Over the past two decades, Argo profiles have provided unprecedented insight into the global patterns of space and time variability of ocean temperature and salinity, significantly reducing associated uncertainties. However, analyzing such assessments during the pre-Argo period remains challenging due to the scarcity of observations in many regions. From the Argo period, a set of dominant three-dimensional patterns can be estimated using empirical orthogonal function (EOF) analysis, which helps to fill in observational gaps. From the associated principal components, temporal fluctuations can be observed, aiming to build a catalog of possible ocean state trajectories. To map pre-Argo observations, EOFs are used in a data assimilation framework that uses this catalog to feed an analog prediction and provide reanalysis. In this study, a new data-driven interpolation method called Reduced-Space Analog Data Assimilation (RedAnDA) 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 subsequently applied to a real historical dataset and compared with other available reanalysis products. Overall, the reconstructed temperature field showed variability consistent with the OSSE true field 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, observational sampling, and unresolved scale issues.

KEYWORDS: ENSO; Empirical orthogonal functions; Statistical forecasting; Data assimilation; Interpolation schemes; Dimensionality reduction

1. Introduction

In the ocean, temperature plays a fundamental role in influencing dynamic processes such as three-dimensional geostrophic circulation and the mixing of water masses, primarily by setting the density in most tropical to subpolar regions. It has a significant effect on ecosystems (Wilson and Coles 2005). Temporal changes in globally averaged ocean temperature reflect not only long-term global warming through heat uptake (Rhein et al. 2013; Von Schuckmann et al. 2023) but also climate and atmospheric variability through air–sea interactions (Bjerknes 1969). Moreover, the ocean exhibits significant variability ranging from regional to basin scales and from decades to seconds, which can superimpose and interact, increasing uncertainties in the estimation of long-term anthropogenic trends or underlying natural variability (Rhein et al. 2013; Palmer et al. 2019; Johnson and Lyman 2020). Notably, phenomena such as air–sea coupled El Niño–Southern Oscillation (ENSO) have a pronounced impact on surface temperature, particularly in the Pacific Ocean, with global impacts on long-term temperature trends (Roemmich and Gilson 2011; Johnson and Birnbaum 2017). 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, favoring better sampling of the Northern Hemisphere and coastal regions (Abraham et al. 2013). Furthermore, the majority of research expeditions were conducted during the summer months, primarily due to favorable weather conditions for navigation. This lack of in situ observations has impeded the ability to consistently describe regional patterns and understand underlying oceanic mechanisms (Gille 2008), contributing to uncertainties in quantifying the rate of ocean heat content change (Freeland et al. 2010; Roemmich et al. 2012; von Schuckmann et al. 2014; Roemmich et al. 2015; Riser et al. 2016).

Since the late 1990s, the Argo global observing system has been deployed, achieving its targeted quasiglobal coverage of $3^\circ \times 3^\circ$ temperature and salinity profiles over the upper 2000 m of the ocean in 2007 (Roemmich et al. 2012, 2019). Over the past two decades, it has become the backbone of the Global Ocean Observing System (GOOS) and now serves as the primary source of observations for numerous ocean and climate studies (see Johnson et al. 2022, 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 confidence in assessments of ocean heat content, thermal expansion, and changes in freshwater content (Hansen et al. 2011;

Corresponding author: Erwan Oulhen, erwan.oulhen@univ-brest.fr

von Schuckmann et al. 2014; Roemmich et al. 2015; Riser et al. 2016).

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. The 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.

First, 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, these methods 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 onward, with a focus on evaluating observational quality. Kaplan et al. (1997) combined data reduction using empirical orthogonal functions with 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. Another example is the work of Cheng and Zhu (2016), who utilized an ensemble OI approach, incorporating a first guess and a priori covariance derived from phase 5 of the Coupled Model Intercomparison Project (CMIP5), to reconstruct ocean subsurface temperature patterns and the trend from 1940 to 2014.

Second, data assimilation procedures in numerical models allow for physically consistent data interpolation (Pham et al. 1998). These analyses give robust estimates of upper ocean heat content (OHC) and the depth of the tropical mixed layer (Balmaseda et al. 2015). However, this approach has significant computational costs, as it requires the integration of the primitive equations into an ocean general circulation model (OGCM) and is subject to inherent numerical imperfections (Forget et al. 2015).

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., Gray and Riser 2015; Tandeo et al. 2015; Sévellec and Drijfhout 2018; Kuusela and Stein 2018; Brown and Caldeira 2020). Leveraging Argo data, Kuusela and Stein (2018) employed stationary Gaussian process regression to compute temperature fields, estimating covariance parameters through local moving-window maximum-likelihood estimation. They enhanced uncertainty estimates by recognizing the non-Gaussian and nonstationary nature of the temperature field. Sévellec and Drijfhout (2018) introduced probabilistic forecast (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 computational costs compared with OGCMs.

In this study, the analog forecasting method (Lorenz 1969) was combined with ensemble data assimilation techniques, forming the analog data assimilation (AnDA) method, as previously introduced by Tandeo et al. (2015) and Lguensat et al. (2017). This method has already shown promising results in previous oceanic applications, such as the successful reconstruction of 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 the depth, during a period when in situ measurements were lacking. To address 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 were the temporal coefficients of the spatial-temporal signals projected into a reduced set of three-dimensional empirical orthogonal functions (EOFs), which were built using the learning period. Utilizing EOFs offered the advantage of extracting dominant patterns of variability, smoothing the analyzed field, and reducing the dimensionality of the system. Lafarga et al. (2023) have demonstrated the effectiveness of using EOFs for reconstruction purposes in the Pacific. The integration of space reduction and AnDA was termed reduced-space AnDA (RedAnDA).

The tropical Pacific was selected as the test region for implementing the ocean temperature reconstruction. This basin exhibits subsurface OHC variability, predominantly driven by the interannual-to-decadal ENSO climate mode (e.g., Chen and Tung 2014). The ENSO phases redistribute heat throughout the basin, with El Niño events resulting in warming of the upper 100 m and cooling of the subsequent 400 m, whereas La Niña events exhibit the reverse pattern. This well-documented interannual variability has been extensively studied (e.g., Roemmich and Gilson 2011; Roemmich et al. 2015; Johnson and Birnbaum 2017). During the Argo period, which served as the learning period for the catalog, six El Niño and five La Niña events were observed.

The paper is organized as follows. First, 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

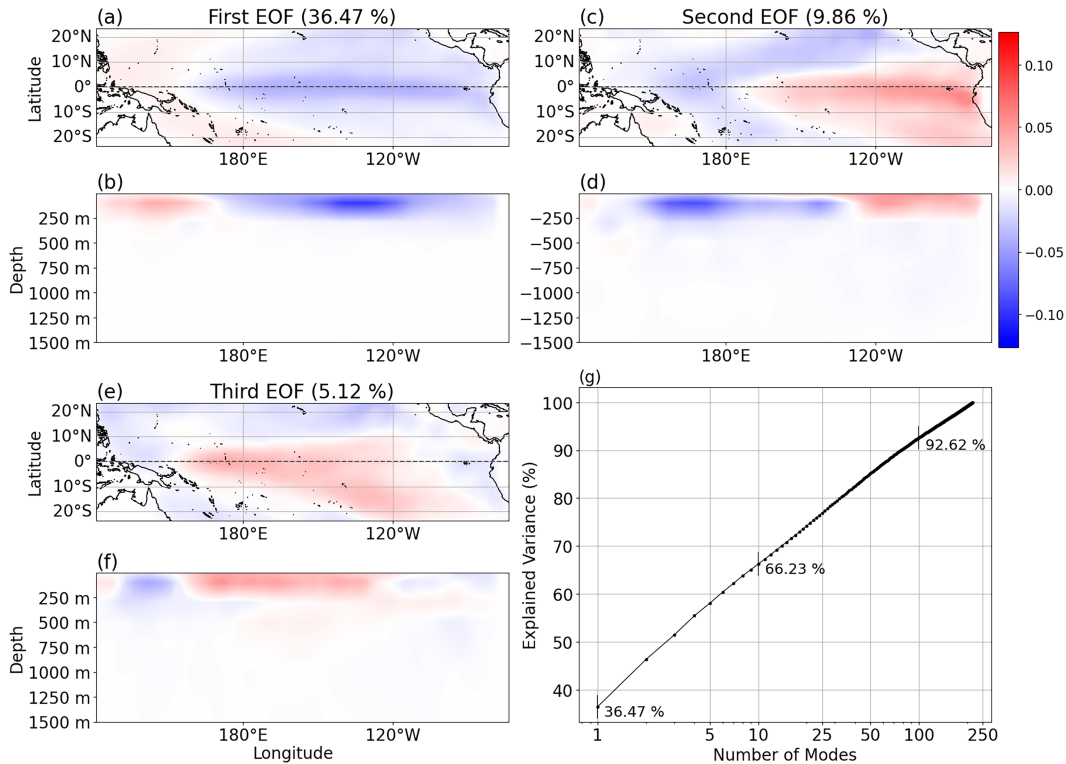


FIG. 1. (a),(b) The first, (c),(d) second, and (e),(f) third EOFs were computed from ISAS20 temperature data at a 5-m depth and along an equatorial section. The explained variance for each of these three EOFs is provided in the titles of (a), (b), and (e). (g) The cumulative explained variance for 252 modes as the number of EOFs increases, with the x axis presented on a log scale. The values of the explained variance for 1, 10, and 100 modes are also indicated.

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 integrates three established methodologies: (i) space reduction utilizing empirical orthogonal functions, (ii) analog prediction, and (iii) data assimilation.

The process began by considering the three-dimensional temperature anomaly field $\mathbf{T}(t)$ from the learning period in the tropical Pacific Ocean, after removing the monthly seasonal cycle. To reduce the dimensionality of the temperature field, EOF decomposition was applied. This technique concentrated the variance into a reduced number of functions, relegating insignificant portions to noise (Emery and Thomson 2001). The reduction in dimensionality not only simplified the system but also facilitated the analog prediction. By retaining the L dominant EOFs, the noise was effectively filtered out, resulting in a relatively small truncation error ϵ' . The selection of L was a critical trade-off between system reduction and the representation of variance. Prior to the decomposition, the temperature field was weighted by multiplying it with $\mathbf{w} \times \mathbf{T}(t)$, where $\mathbf{w}$ represents the square root of the cosine of latitude. This weighting ensured that the variance had a uniform impact

across different regions, accounting for their surface area (Baldwin et al. 2009; Shea 2013).

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

$$\mathbf{w} \times \mathbf{T} = \alpha(t) \text{ EOF} + \epsilon'(t). \quad (1)$$

The EOF decomposition was applied to the tropical Pacific using the learning period. Energy redistribution was performed according to Eq. (19) in Kaplan et al. (1997), to remedy the fact that the variance spectrum of the EOFs decreases too steeply due to the initial coarsening of the field, leading to a strong constraint and low error estimates. The patterns of variability were associated with a tongue in the equatorial band, from the eastern to central Pacific, and the warm pool in the western Pacific (Figs. 1a,b). The transect (Fig. 1b) displayed a concentration of correlation in the subsurface layers, primarily within the first 200 m. These patterns were typical of eastern Pacific El Niño events (Kao and Yu 2009). The second mode captured the information on the variability of central Pacific El Niño events (Kao and Yu 2009) (Figs. 1c,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 was required to express the diversity of El Niño events (Trenberth and Stepaniak 2001). The subsequent modes held lesser physical significance. The third mode was not easily interpreted (Figs. 1e,f). The following modes explained on average 0.23% of the initial variance (Fig. 1g). Nonetheless, they remained valuable for consistently conveying information from observations to unsampled locations, even at depth. A fundamental assumption of this 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 replaced an explicit dynamical model with the analog backward prediction, also known as analog backcast. It temporally propagated the interpolated information, providing informative a priori estimates of the ocean state, which were subsequently updated using available observations at corresponding times.

The use of analog prediction as a dynamical model offered several advantages, including eliminating the need to run ocean general circulation models and reducing computational costs. However, it required a dataset representative of the system's dynamics, ideally extensive enough to encompass all potential system states. This dataset, termed a catalog, included 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 could be constructed from observational data or numerical simulation results. In this study, the analogs consisted of temporal coefficients $\alpha(t)$ and their predecessors $\alpha(t - dt)$ 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:

$$g(t, t') = \exp[-\delta^2 |\alpha(t) - \alpha(t')|^2], \quad (2)$$

where t represents the time when the backcast is performed and t' is another time for which a record exists in the catalog, the Euclidean distance is denoted as $|\cdot|$, and δ is set as the inverse of 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 the 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, increased, the search for analogs rapidly became more difficult, a phenomenon known as the “curse of dimensionality” (Clarkson 1994). This complexity made the reduced-space approach convenient in this case.

At time t , when the analog backcast $\mathcal{A}$ was applied to obtain an estimate of $\alpha(t - dt)$, the k analogs of the current state $\alpha(t)$ and their predecessors were selected. Each pair of analog and predecessor was then weighted according to Eq. (2). Assuming that both selections belonged to normal distributions,

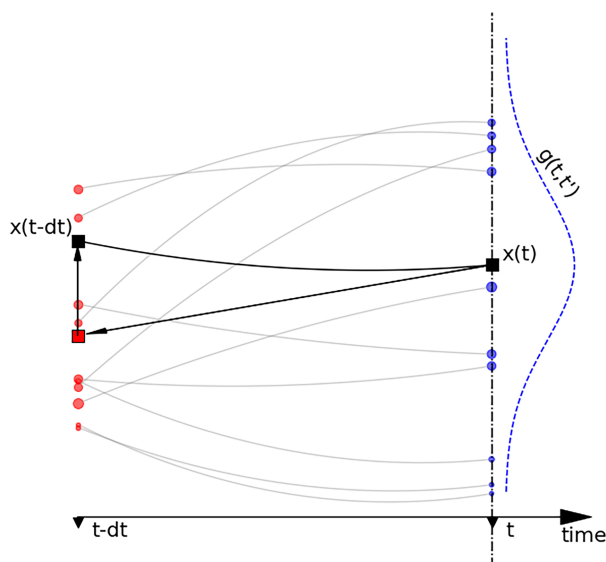


FIG. 2. The process of analog backcast prediction. The black square $x(t)$ represents the current state for which backward evolution is computed. The blue dots represent the analogs, and the red dots represent their temporal predecessors. In this context, the size of the analog sample is $k = 10$. The size of each dot reflects the weighting according to similarity, as determined by the function $g(t, t')$ from Eq. (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.

a linear fit was performed, following the weighted least squares method of Cleveland (1979). The obtained linear regression was used to extend $\alpha(t)$ at $t - dt$, providing a first estimate of the backcast (see Fig. 2). The final backcast was 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 reflected the uncertainties in the prediction of the analog method, which were partially controlled by the chaotic dynamics of the system. When it was observed that the linear fit efficiently linked the most important analogs and their associated predecessors, the prediction was 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 were reflected by changes in the persistence of α samples. This adaptive approach was an advantage over sequential interpolation techniques that determined temporal persistence mostly on a priori statistics. The uncertainty of the analog prediction was 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 in situ analysis system (ISAS20) temperature field (Gaillard et al. 2016). Associated $\alpha(t)$ values between 2002 and 2020 were used in the catalog of RedAnDA.

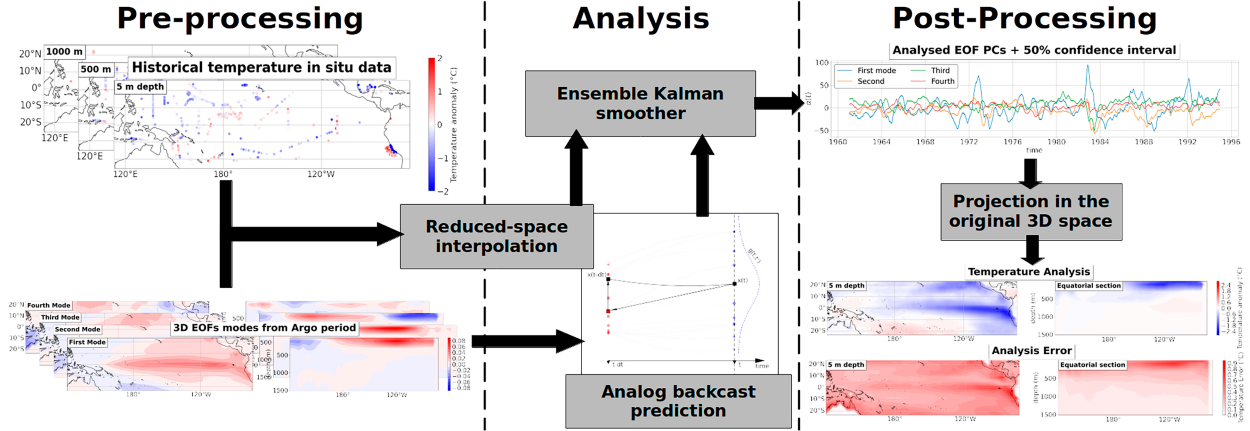


FIG. 3. (center) RedAnDA analysis, initiated by an initial temperature product drawn from a (left) learning period and preprocessing of observation data. (right) The postanalysis steps lead to the final gridded temperature product.

The reduced set of $\alpha(t)$ 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 the EN.4.2.2 dataset (Good et al. 2013).

In this study, sequential RedAnDA was executed in a backward temporal fashion. By starting from the state $\alpha(t + dt)$, $\alpha(t)$ was determined through the application of the analog backcast $\mathcal{A}$. A notable distinction from the forward-time analysis was observed during the initialization phase, wherein the initial α 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:

$$\alpha(t) = \mathcal{A}[\alpha(t + dt), \epsilon^m(t + dt)], \quad (3)$$

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

The analog backcast $\mathcal{A}$ uncertainty is denoted as ϵ^m . The observed data are represented by $\mathbf{y}$, while $\mathcal{H}(t)$ 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 $\epsilon^o(t)$.

The reduced-space interpolation method proposed by Kaplan et al. (1997) was consequently employed to define $\mathcal{H}(t)$ and $\epsilon^o(t)$, with detailed elaboration provided later in the section.

The analysis proceeded in accordance with the ensemble Kalman smoother, augmented with the analog backcast as depicted in Fig. 1 of Tandeo et al. (2015). Within the ensemble variant of the Kalman filter, $\alpha(t)$ and $\mathbf{y}(t)$ were treated as stochastic variables, following normal law distributions whose mean value and variance were determined via a Monte Carlo procedure. The variance served to estimate the uncertainty associated with the variable.

Once $\alpha(t)$ has been reconstructed for all desired time steps, the spatiotemporal temperature anomaly field was subsequently derived through reprojection into real space, utilizing the EOF patterns. The entire RedAnDA procedure is summarized in Fig. 3. The ensemble Kalman filter process is schematized in Fig. 4. 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 was assessed through an OSSE (Masutani et al. 2010). The method was applied to the monthly output from the Oceanic Chaos-Impacts, Structure, Predictability (OCCIPUT) project (Penduff et al. 2014; Bessi eres et al. 2017). The simulation spanned the period from 1961 to 2015. The 408 months between 1961 and 1994 were 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 were available, provided analogs and predecessors for the analog prediction. The full grid maps of the reconstruction period were further used to qualify the target truth. The way these periods were split was designed to mimic the real situation where a recent and short Argo period had good sampling and a longer pre-Argo period had fewer observations.

In the OSSE, for comparison with the RedAnDA method, a standard optimal interpolation method was also implemented, following Good et al. (2013). This method reconstructed monthly temperature maps sequentially, from 1961 to 1994, using an a priori monthly climatology built during the learning period, to which approximately 82% of the previous month's analyzed anomaly was added.

To reduce the calculations and focus on the large spatial scale, all gridded products, for the OSSE and for real reconstruction, were coarsened down to $3^\circ \times 3^\circ$ by horizontal averaging. Depth levels were restricted to 5, 100, 300, 500, 700, 1000, and 1500 m. The tropical Pacific was defined in this study as the region between 120°E and 70°W, and 25°S and

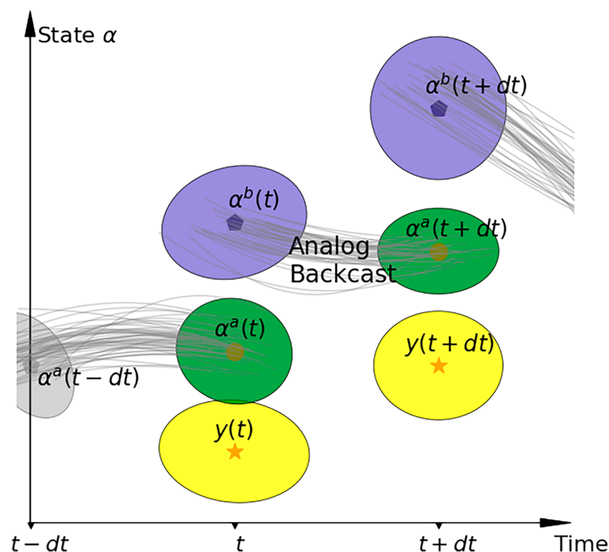


FIG. 4. Schematic diagram of the Kalman filter process within the data assimilation method. Beginning with the analyzed state α^a resulting from the assimilation of observations (depicted in green), the analog backcast was executed to provide an estimate of the state at the previous time step α^b (represented 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).

25°N. EOFs were 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 at 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 m were reconstructed by approximately 67%, 73%, 44%, 27%, 22%, 23%, and 20%, respectively.

In practice, $\epsilon^o(t)$ from Eq. (4) was computed by combining ϵ^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 EN.4.2.2 profiles found in large horizontal boxes, at each depth. For the real observation uncertainty, $\epsilon^o(t)$ also encompassed instrumental error, as assessed following Cowley et al. (2021). This error ranged from 0.3°C for mechanical bathythermograph (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

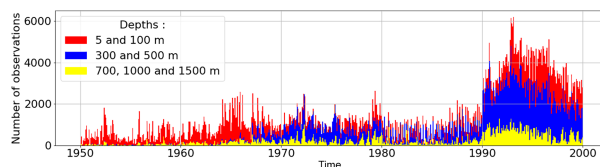


FIG. 5. Monthly histogram of available EN.4.2.2 temperature observations in the tropical Pacific, categorized by depth, for the period between 1950 and 2000. In red color, observations at 5- and 100-m depth are counted. In blue color, observations at 300- and 500-m depth are counted. The number of observations at 700-, 1000-, and 1500-m depth is represented in yellow.

minimum of 5000 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 gridded product was selected to represent the Argo period (Gaillard et al. 2016). This product provides monthly $0.5^\circ \times 0.5^\circ$ gridded temperature and salinity fields, constructed using an optimal interpolation method. It encompasses 187 standard depth levels distributed between 0 and 5500 m 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 correlation scales were used: 300 km and four times the deformation radius (Gaillard et al. 2016).

The assimilated profiles for reconstructing past periods came from the EN.4.2.2 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 Programme (GTSP) (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 EN.4.2.2, only depths where the temperature observation had a quality flag of 1 were 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), the time series were averaged over 15-day periods for each identified buoy. Superprofiles were then considered at the center of the time slice, reducing the number of initial profiles from 832 703 to 648 786. In Fig. 5, the number of available observations, between 1950 and 2000, appears to increase over time, with a substantial improvement in sampling in the 1990s. Two minima were observed in the 1950s

and 1980s, suggesting that reconstruction during these periods may have been less constrained.

In the OSSE framework, the OCCIPUT project was utilized (Penduff et al. 2014; Bessi eres et al. 2017). The OCCIPUT project offers an ensemble of 50 realizations of a global $1/4^\circ$ ocean/sea ice simulation, based on the NEMO 3.5 model (Penduff et al. 2014; Bessi eres et al. 2017), alongside synthetic profiles extracted at specific times and locations corresponding to the historical database. For this study, only one member of the model output and synthetic profiles was used. The simulation spans the period from 1961 to 2015, employing the Drakkar forcing set (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 were 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 utilized in situ observations from the WOD05, the GTSP, Centennial In Situ Observation Based Estimates (COBE), and Argo buoys (Ishii et al. 2005). It compiled monthly $1^\circ \times 1^\circ$ temperature fields on 24 depth levels.
- EN4: A gridded product from the Hadley Centre of the Met Office. EN4 employed 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 provided monthly temperature fields from 1900 to the present day over 42 depth levels.
- IAP product: Provided by ensemble optimal interpolation of the CAS-Ocean Data Center, Global Ocean Science Database (CODC-GOSD), it utilized CMIP5 models to reconstruct global and gridded temperature at a monthly period from 1940 to the present day. Spatial resolution was $1^\circ \times 1^\circ$ in the upper 2000 m, over 41 levels (Cheng and Zhu 2016; Cheng et al. 2017).
- Estimated State of Ocean for Climate Research (ESTOC) dataset: Developed by the Japan Agency for Marine-Earth Science and Technology (JAMSTEC) and Kyoto University, ESTOC utilized a four-dimensional variational data assimilation method with the GFDL Modular Ocean Model, version 3 (MOM3). It assimilated quality-controlled EN3 in situ profiles and sea surface dynamic height data from AVISO altimetry products (Osafune et al. 2014, 2015). The dataset provides a 1×1 monthly result covering the period from 1956 to 2009 covering 46 vertical levels.
- Ocean Reanalysis System 4 (ORAS4): Provided by the Copernicus Climate Change Service, ORAS4 combined model data with observations using the NEMO model and the 3D-Var assimilation system called NEMOVAR (Balmaseda et al. 2013). It assimilated in situ profiles from the EN3 dataset, sea surface temperature from the OI version 2 (OIv2) product, and altimeter sea level anomalies from the AVISO dataset. ORAS4 covers the period from 1959 to the present

day over 42 depth levels with a horizontal spatial resolution of $1^\circ \times 1^\circ$.

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 OCCIPUT model fields from 1995 to 2015 (Penduff et al. 2014; Bessi eres et al. 2017) served as a learning period for the analog backcast.

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

The reconstructions in the tropical Pacific at the 100-m depth level (Figs. 6a–c), during the 1973 El Ni o, as identified in the Ni o-3.4 time series (Fig. 6d), showed that the temperature field was consistently reconstructed with RedAnDA (Figs. 6a,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 Ni o 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 $5^\circ\text{--}10^\circ\text{N}$ was satisfactorily reconstructed with RedAnDA (Figs. 6a,c). Due to an insufficient number of observations and inaccurate a priori correlation scales, OI did not reproduce such fine structure (Figs. 6b,c).

The time series of the average temperature anomaly in the Ni o-3.4 box, a region where an SST anomaly index is typically calculated, indicated that both methods performed better in the recent period, from 1974 to 1994, when more observations were available (Figs. 5 and 6d). In this selected region, the temperature anomaly was strongly positive during El Ni o 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 Ni o events of 1963, 1965, 1972, and 1976 were systematically underestimated by about 1°C by the OI reconstruction. Similarly, most La Ni a events were identified in both products (1962, 1964, 1971, 1973, 1974, 1975, 1984, and 1988), but tended to be overestimated by $1^\circ\text{--}2^\circ\text{C}$, 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 (Figs. 7a–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 (Figs. 7a–c). Although RedAnDA and OI reconstructed the eastern warm and northwestern cold regions, they did so with an

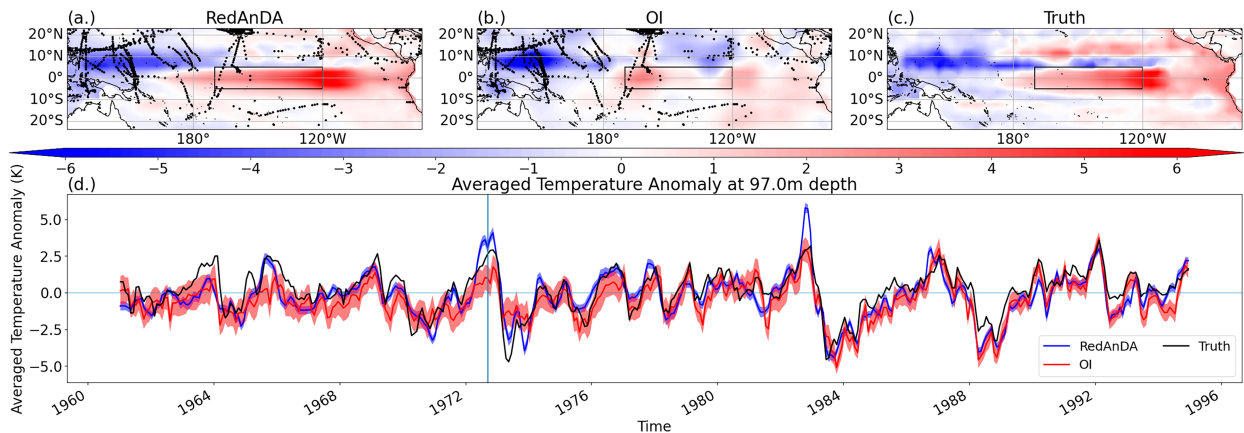


FIG. 6. Temperature anomalies at around 100-m depth for (a) RedAnDA, (b) OI, and (c) OSSE truth in September 1972. The black dots indicate synthetic observations available at the given depth and time. The black box in (a)–(c) shows the region where the temperature was calculated to obtain a pseudo-Niño-3.4 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.

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.

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 (Figs. 7c,d). This specific region was selected 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 (Li et al. 2020). 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 1968–74 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.

At around 1000-m depth, the number of observations diminished significantly (Figs. 5 and 8). This 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 (Figs. 8b,c). In

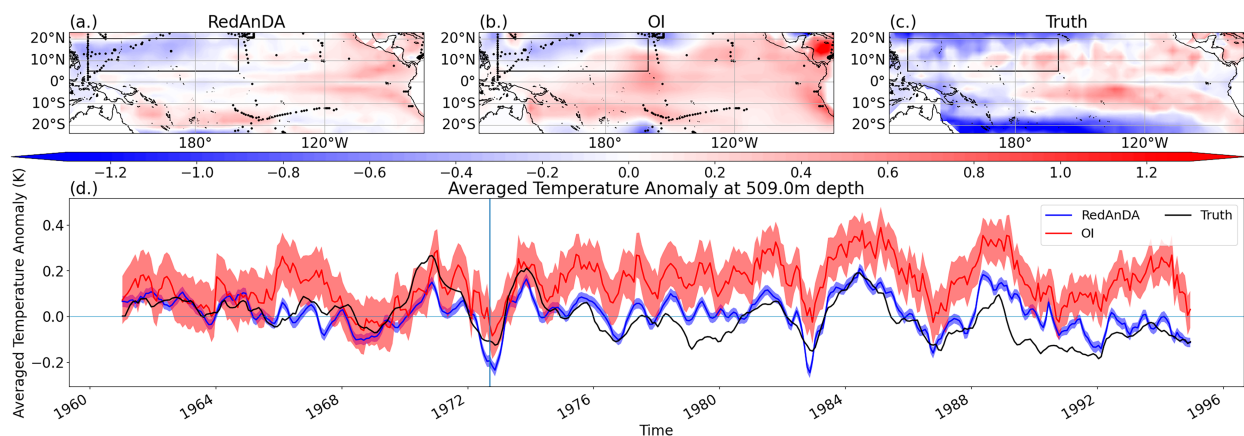


FIG. 7. Temperature anomalies around 500-m depth for (a) RedAnDA, (b) OI, and (c) OSSE truth in September 1972. Black dots indicate synthetic observations available at the given depth and time. The black box in (a)–(c), now located between 5° – 20°N and 130°E – 160°W , shows the region where the temperature was calculated to estimate the signature of the westward NESC (Li et al. 2020). (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.

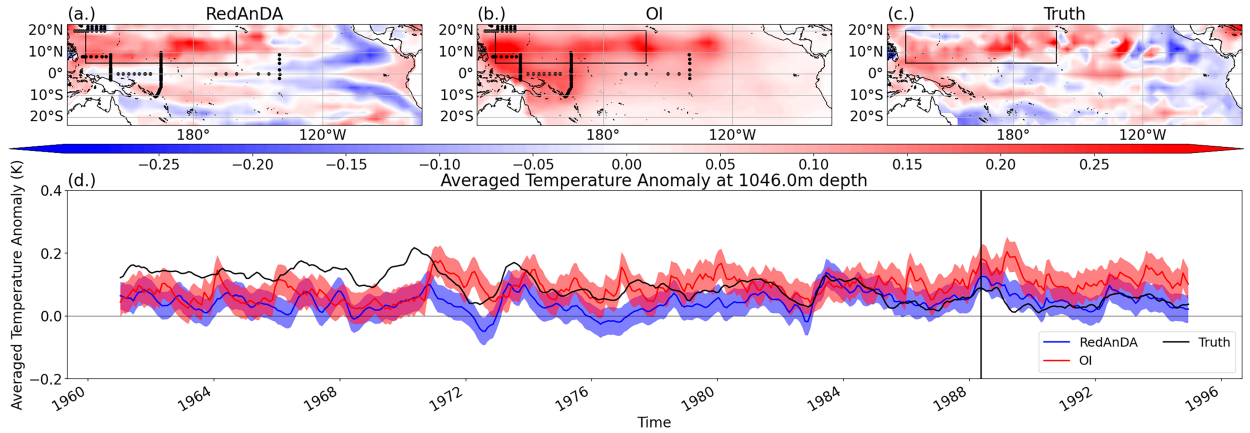


FIG. 8. Temperature anomalies around 1000-m depth for (a) RedAnDA, (b) OI, and (c) OSSE truth in May 1988. Black dots indicate synthetic observations available at the given depth and time. The black box in (a)–(c), between 5°–20°N and 130°E–160°W, shows the region where the temperature was averaged to assess the variability associated with the north equatorial deep jets (Cravatte et al. 2012; Qiu et al. 2013; Wang et al. 2020). (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.

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 and 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 (Figs. 8a,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, probability coverage was computed. It was 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 Tandeo et al. (2020). The probability coverage is given by

$$\text{probability coverage} = P[\mathbf{T}(t) - q \times \boldsymbol{\sigma} \geq \mathbf{T}^{\text{true}}(t) \geq \mathbf{T}(t) + q \times \boldsymbol{\sigma}]. \quad (5)$$

Generally, both methods captured less than the desired amount of truth within their confidence intervals (Figs. 9e–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 domains where a priori variances were the highest (Figs. 9c,d,g,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% and 40% (Fig. 9g).

The uncertainty in the RedAnDA method was influenced by factors such as spatial reduction, the sampling of observations, and the dispersion of ensemble members during back-cast predictions. 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 the drift was significant (Fig. 8d). Overall, the results were overly confident at the surface, as the inclusion of the truth within the confidence interval was less than 30%. The quantification of uncertainty improved with depth (Figs. 9e,f).

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

$$\text{variance ratio} = \frac{\overline{\mathbf{T}(t)^2}}{\overline{\mathbf{T}^{\text{true}}(t)^2}}. \quad (6)$$

The variance ratio compared the variance of the reconstructed signal $\mathbf{T}$ to the variance of the true signal $\mathbf{T}^{\text{true}}$. A

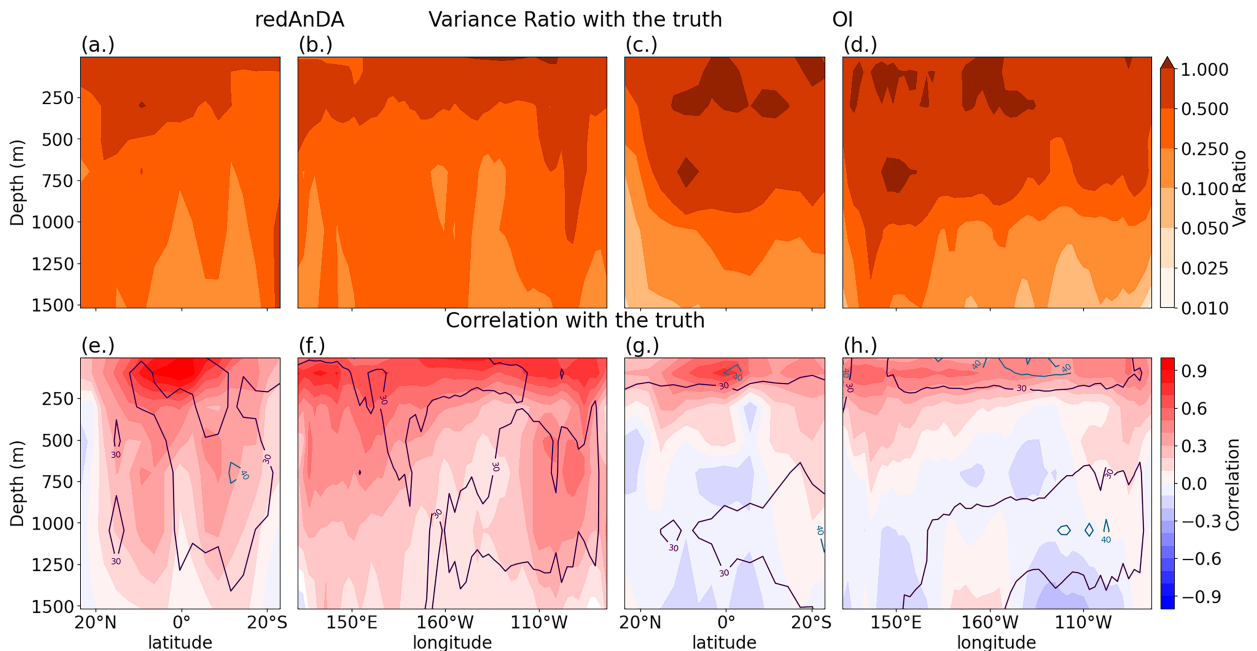


FIG. 9. Meridional- and zonal-mean sections depicting (a)–(d) the ratio of reconstructed variance to true variance and (e)–(h) the correlation of the two analysis products with the OSSE truth. The purple and blue contours in (e)–(h) indicate the 30% and 40% values in probability coverage. For the variance ratio, note the logarithmic color scale. The RedAnDA products are shown in (a), (b), (e), and (f); the OI products are shown in (c), (d), (g), and (h).

variance ratio of 1 indicated that the reconstructed variability matched the true variability in energy:

$$\text{correlation} = \frac{[\mathbf{T}(t) - \mathbf{T}^{\text{true}}(t)]^2}{\sqrt{\mathbf{T}(t)^2 \times \mathbf{T}^{\text{true}}(t)^2}}. \quad (7)$$

The correlation measured the coherence between the true and reconstructed signals. It ranged from -1 to 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 reproduce a satisfactory level of variability in the upper 300 m, where they also exhibited satisfactory correlations with the model truth (Figs. 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 (Figs. 9a,b,e,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 (Figs. 9c,d,g,h).

At depths exceeding 300 m, OI continued to inject information but displayed weak or negative correlations (ranging from 0 to -0.2) with the ground truth (Figs. 9c,d,g,h). The signal reconstructed by RedAnDA exhibited more consistency with the truth, displaying positive correlation reaching up to 0.5 (Figs. 9a,b,e,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.

First, SST reconstruction maps were examined for the intense 1982 El Niño event (Figs. 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° and 150°W , the cold anomaly was detected by all methods. 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. 10a).

The Niño-3.4 index exhibited good agreement among all time series (Fig. 10g). The 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.

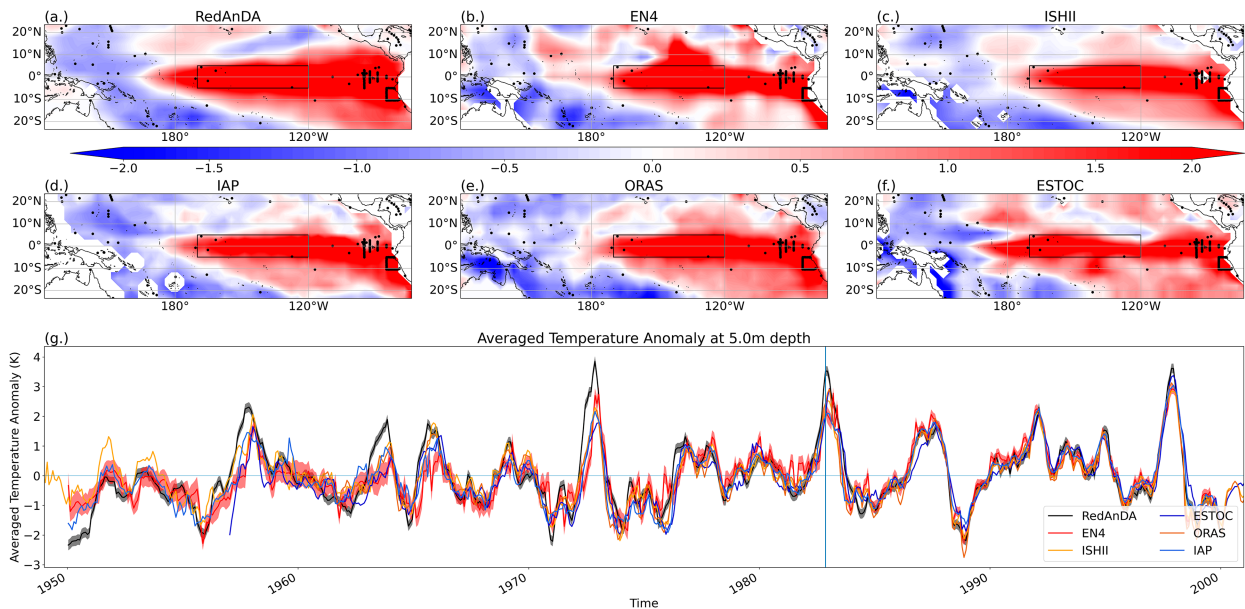


FIG. 10. Temperature anomalies at 5-m depth in November 1982 for various datasets, including (a) RedAnDA, (b) EN4, (c) ISHII, (d) IAP, (e) ORAS, and (f) ESTOC. In (a)–(f), snapshots of temperature anomalies for each dataset are shown, with black dots indicating available observations at the given depth 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. (g) 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.

At a depth of 300 m, in February 1978, the various methods yielded less consistent results (Figs. 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 warm anomalies ranging between 0.5° and 1°C on its southern and northern flanks. 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 (Figs. 11a,d,e,f). In this region, RedAnDA diverged from EN4 and ISHII, but these reconstructions, lacking sufficient observational data, exhibited low energy (Figs. 11b,c).

In the box (Figs. 11a–f), located in the southwest of the study area, between 150°E–160°W and 5°–15°S, temporal fluctuations were related in part to the Pacific South Equatorial Current (SEC). This current undergoes interannual fluctuations in its intensity related to ENSO, flowing stronger a few months following an El Niño event and weaker after a La Niña event (Kessler and Cravatte 2013). Regarding the estimation of this variability, after the 1970s, the different time series are aligned (Fig. 11g). 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 (Figs. 11a–f).

A similar discrepancy occurred in 1959. In the 1950s, 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 other products indicated overall agreement (Table 1). ISHII and ORAS showed the closest similarity to RedAnDA, with correlations of 0.61, while ESTOC showed the least similarity with a value of 0.47. These close scores demonstrated that RedAnDA aligned with 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 upper 500-m depth may be worth consideration as it may have reconstructed components of the variability that were undetected by other datasets.

6. Discussion and conclusions

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

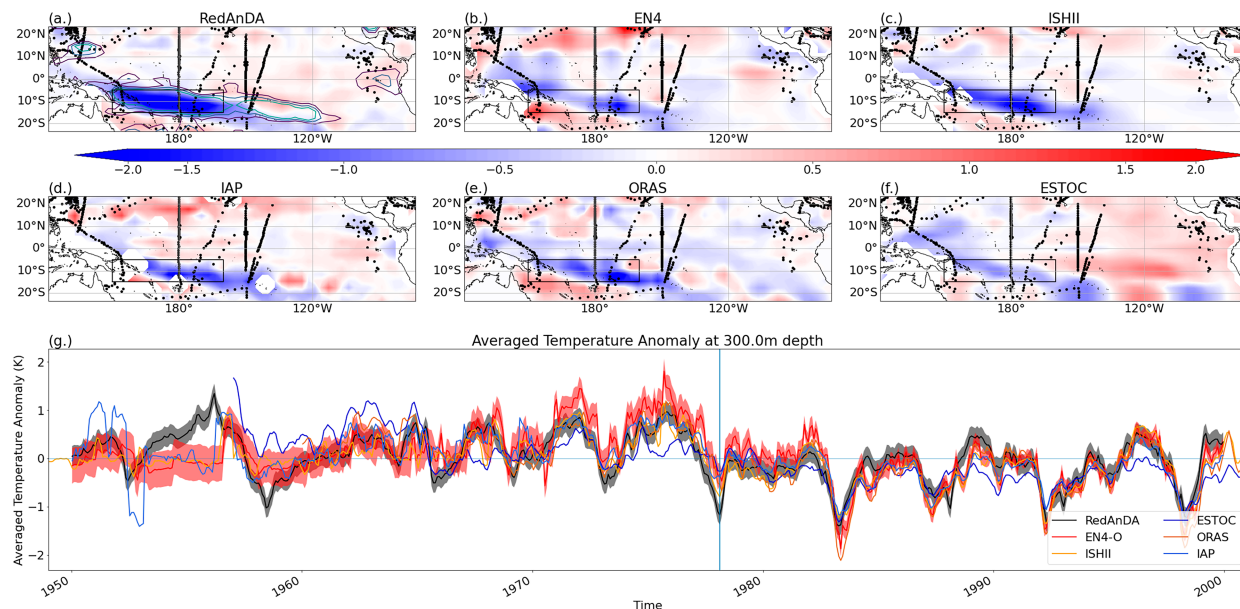


FIG. 11. Temperature anomalies at 300-m depth in February 1978, including (a) RedAnDA, (b) EN4, (c) ISHII, (d) IAP, (e) ORAS, and (f) ESTOC. In (a)–(f), snapshots of temperature anomalies for each dataset are shown, with black dots indicating available observations at the given depth and time. The observation 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 SEC (Wyrski 1974; Ganachaud et al. 2014; Yang et al. 2022). (g) 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.

optimal 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 provided statistical temporal predictions that enhanced the estimation of the a priori statistics at 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 nonstationarity 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 (Clarkson 1994). 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.

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.

	Correlation
ISHII	0.61
EN4	0.52
IAP	0.57
ORAS	0.61
ESTOC	0.47

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 (Stammer 2005; Forget et al. 2015). Given that the influence of such factors was not assessed in this case, it was decided not to incorporate simulated data into this reconstruction of reality.

Within the OSSE framework, it was envisaged to explore a broader catalog by incorporating additional members from the OCCIPUT ensemble (Penduff et al. 2014). However, 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 (Sérazin et al. 2017). 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 ISAS20, 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 ISAS20 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 twentieth century, notably at depth, may be characterized through the use of the new temperature product provided by RedAnDA.

Reconstructing temperature and salinity with RedAnDA in a covarying approach could provide new datasets of the evolution of regional water masses over the second half of the twentieth century. Considering their covariability would be beneficial to salinity due to 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 precursor method (e.g., Moore and Kleeman 1999, 2001; Moore et al. 2003; Mu et al. 2013), the analogs could provide new means of identifying conditions conducive to the 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 <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 at <https://doi.org/10.17882/52367>; EN.4.2.2 profiles and objective analyses are available at <https://www.metoffice.gov.uk/hadobs/en4/download-en4-2-2.html>; OCCIPUT data are available upon demand; ORAS4 data are available at https://icdc.cen.uni-hamburg.de/thredds/dodsC/oras4_temp_all.html; IAP data are available at http://www.ocean.iap.ac.cn/ftp/cheng/CZ16_v3_IAP_Temperature_gridded_1month_netcdf/; ISHII data are available at <https://rda.ucar.edu/datasets/ds285.3/dataaccess/>; and ESTOC data are available at <https://www.godac.jamstec.go.jp/jagdas/catalog/estoc/02c/catalog.html>. RedAnDA results and codes will be provided upon demand.

REFERENCES

- Abraham, J. P., and Coauthors, 2013: A review of global ocean temperature observations: Implications for ocean heat content estimates and climate change. *Rev. Geophys.*, **51**, 450–483, <https://doi.org/10.1002/rog.20022>.
- Baldwin, M. P., D. B. Stephenson, and I. T. Jolliffe, 2009: Spatial weighting and iterative projection methods for EOFs. *J. Climate*, **22**, 234–243, <https://doi.org/10.1175/2008JCLI2147.1>.
- Balmaseda, M. A., K. Mogensen, and A. T. Weaver, 2013: Evaluation of the ECMWF ocean reanalysis system ORAS4. *Quart. J. Roy. Meteor. Soc.*, **139**, 1132–1161, <https://doi.org/10.1002/qj.2063>.
- , and Coauthors, 2015: The Ocean Reanalyses Intercomparison Project (ORA-IP). *J. Oper. Oceanogr.*, **8**, s80–s97, <https://doi.org/10.1080/1755876X.2015.1022329>.
- Bessi eres, L., and Coauthors, 2017: Development of a probabilistic ocean modelling system based on NEMO 3.5: Application at eddying resolution. *Geosci. Model Dev.*, **10**, 1091–1106, <https://doi.org/10.5194/gmd-10-1091-2017>.
- Bjerknes, J., 1969: Atmospheric teleconnections from the equatorial Pacific. *Mon. Wea. Rev.*, **97**, 163–172, [https://doi.org/10.1175/1520-0493\(1969\)097<0163:ATFTEP>2.3.CO;2](https://doi.org/10.1175/1520-0493(1969)097<0163:ATFTEP>2.3.CO;2).
- Boyer, T. P., and Coauthors, 2006: World Ocean Database 2005. S. Levitus, Ed., NOAA Atlas NESDIS 60, 190 pp., <https://repository.library.noaa.gov/view/noaa/1131>.
- Bretherton, F. P., R. E. Davis, and C. B. Fandry, 1976: A technique for objective analysis and design of oceanographic experiments applied to MODE-73. *Deep-Sea Res. Oceanogr. Abstr.*, **23**, 559–582, [https://doi.org/10.1016/0011-7471\(76\)90001-2](https://doi.org/10.1016/0011-7471(76)90001-2).
- Brown, P. T., and K. Caldeira, 2020: Empirical prediction of short-term annual global temperature variability. *Earth Space Sci.*, **7**, e2020EA001116, <https://doi.org/10.1029/2020EA001116>.

- Chen, X., and K.-K. Tung, 2014: Varying planetary heat sink led to global-warming slowdown and acceleration. *Science*, **345**, 897–903, <https://doi.org/10.1126/science.1254937>.
- Cheng, L., and J. Zhu, 2016: Benefits of CMIP5 multimodel ensemble in reconstructing historical ocean subsurface temperature variations. *J. Climate*, **29**, 5393–5416, <https://doi.org/10.1175/JCLI-D-15-0730.1>.
- , K. E. Trenberth, J. Fasullo, T. Boyer, J. Abraham, and J. Zhu, 2017: Improved estimates of ocean heat content from 1960 to 2015. *Sci. Adv.*, **3**, e1601545, <https://doi.org/10.1126/sciadv.1601545>.
- Clarkson, K. L., 1994: An algorithm for approximate closest-point queries. *SCG'94: Proc. 10th Annual Symp. on Computational Geometry*, Stony Brook, NY, Association for Computing Machinery, 160–164, <https://doi.org/10.1145/177424.177609>.
- Cleveland, W. S., 1979: Robust locally weighted regression and smoothing scatterplots. *J. Amer. Stat. Assoc.*, **74**, 829–836, <https://doi.org/10.1080/01621459.1979.10481038>.
- Cowley, R., and Coauthors, 2021: International Quality-Controlled Ocean Database (IQuOD) v0.1: The temperature uncertainty specification. *Front. Mar. Sci.*, **8**, 689695, <https://doi.org/10.3389/fmars.2021.689695>.
- Cravatte, S., W. S. Kessler, and F. Marin, 2012: Intermediate zonal jets in the tropical Pacific Ocean observed by Argo floats. *J. Phys. Oceanogr.*, **42**, 1475–1485, <https://doi.org/10.1175/JPO-D-11-0206.1>.
- Dussin, R., B. Barnier, L. Brodeau, and J. M. Molines, 2016: The making of the Drakkar forcing set DFS5. MyOcean Rep. 01-04-16, 34 pp., https://www.drakkar-ocean.eu/publications/reports/report_DFS5v3_April2016.pdf.
- Emery, W. J., and R. E. Thomson, 2001: The spatial analyses of data fields. *Data Analysis Methods in Physical Oceanography*, W. J. Emery and R. E. Thomson, Eds., Elsevier Science, 305–370, <https://doi.org/10.1016/B978-044450756-3/50005-8>.
- Forget, G., J.-M. Campin, P. Heimbach, C. N. Hill, R. M. Ponte, and C. Wunsch, 2015: ECCO version 4: An integrated framework for non-linear inverse modeling and global ocean state estimation. *Geosci. Model Dev.*, **8**, 3071–3104, <https://doi.org/10.5194/gmd-8-3071-2015>.
- Freeland, H. J., and Coauthors, 2010: Argo—A decade of progress. *Proc. OceanObs'09: Sustained Ocean Observations and Information for Society*, Vol. 2, ESA Special Publication WPP-306, Venice, Italy, European Space Agency, 357–370, <https://archimer.ifremer.fr/doc/00029/14038/>.
- 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. *J. Climate*, **29**, 1305–1323, <https://doi.org/10.1175/JCLI-D-15-0028.1>.
- Ganachaud, A., and Coauthors, 2014: The Southwest Pacific Ocean Circulation and Climate Experiment (SPICE). *J. Geophys. Res. Oceans*, **119**, 7660–7686, <https://doi.org/10.1002/2013JC009678>.
- Gille, S. T., 2008: Decadal-scale temperature trends in the Southern Hemisphere ocean. *J. Climate*, **21**, 4749–4765, <https://doi.org/10.1175/2008JCLI2131.1>.
- 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. *J. Geophys. Res. Oceans*, **118**, 6704–6716, <https://doi.org/10.1002/2013JC009067>.
- Gray, A. R., and S. C. Riser, 2015: A method for multiscale optimal analysis with application to Argo data. *J. Geophys. Res. Oceans*, **120**, 4340–4356, <https://doi.org/10.1002/2014JC010208>.
- 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>.
- Ishii, M., M. Kimoto, K. Sakamoto, and S.-I. Iwasaki, 2005: Sub-surface temperature and salinity analyses. National Center for Atmospheric Research, Computational and Information Systems Laboratory, accessed 18 April 2018, <https://doi.org/10.5065/Y6CR-KW66>.
- Johnson, G. C., and A. N. Birnbaum, 2017: As El Niño builds, Pacific warm pool expands, ocean gains more heat. *Geophys. Res. Lett.*, **44**, 438–445, <https://doi.org/10.1002/2016GL071767>.
- , and J. M. Lyman, 2020: Warming trends increasingly dominate global ocean. *Nat. Climate Change*, **10**, 757–761, <https://doi.org/10.1038/s41558-020-0822-0>.
- , and Coauthors, 2022: Argo—Two decades: Global oceanography, revolutionized. *Annu. Rev. Mar. Sci.*, **14**, 379–403, <https://doi.org/10.1146/annurev-marine-022521-102008>.
- Kao, H.-Y., and J.-Y. Yu, 2009: Contrasting eastern-Pacific and central-Pacific types of ENSO. *J. Climate*, **22**, 615–632, <https://doi.org/10.1175/2008JCLI2309.1>.
- Kaplan, A., Y. Kushnir, M. A. Cane, and M. B. Blumenthal, 1997: Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures. *J. Geophys. Res.*, **102**, 27 835–27 860, <https://doi.org/10.1029/97JC01734>.
- Kessler, W. S., and S. Cravatte, 2013: ENSO and short-term variability of the South Equatorial Current entering the Coral Sea. *J. Phys. Oceanogr.*, **43**, 956–969, <https://doi.org/10.1175/JPO-D-12-0113.1>.
- Kolodziejczyk, N., A. Prigent-Mazella, and F. Gaillard, 2021: ISAS temperature and salinity gridded fields. Accessed 20 July 2022, <https://doi.org/10.17882/52367>.
- Kuusela, M., and M. L. Stein, 2018: Locally stationary spatio-temporal interpolation of Argo profiling float data. *Proc. Roy. Soc.*, **474A**, 20180400, <https://doi.org/10.1098/rspa.2018.0400>.
- Lafarga, D., T. Bui, Y. T. Song, T. M. Smith, and S. S. P. Shen, 2023: A feasibility study of three-dimensional empirical orthogonal functions from the NASA JPL ocean general circulation model: Computing, visualization and interpretation. *Tellus*, **75A**, 213–230, <https://doi.org/10.16993/tellusa.3223>.
- Lguensat, R., P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet, 2017: The analog data assimilation. *Mon. Wea. Rev.*, **145**, 4093–4107, <https://doi.org/10.1175/MWR-D-16-0441.1>.
- Li, X., Y. Yang, R. Li, L. Zhang, and D. Yuan, 2020: Structure and dynamics of the Pacific North equatorial subsurface current. *Sci. Rep.*, **10**, 11758, <https://doi.org/10.1038/s41598-020-68605-y>.
- Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646, [https://doi.org/10.1175/1520-0469\(1969\)26<636:APARBN>2.0.CO;2](https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2).
- Masutani, M., and Coauthors, 2010: Observing system simulation experiments. *Data Assimilation*, W. Lahoz, B. Khattatov, and R. Menard, Eds., Springer, 647–679, https://doi.org/10.1007/978-3-540-74703-1_24.
- Moore, A. M., and R. Kleeman, 1999: The nonnormal nature of El Niño and intraseasonal variability. *J. Climate*, **12**, 2965–2982, [https://doi.org/10.1175/1520-0442\(1999\)012<2965:TNNOEN>2.0.CO;2](https://doi.org/10.1175/1520-0442(1999)012<2965:TNNOEN>2.0.CO;2).

- , and —, 2001: The differences between the optimal perturbations of coupled model of ENSO. *J. Climate*, **14**, 138–163, [https://doi.org/10.1175/1520-0442\(2001\)014<0138:TDBTOP>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014<0138:TDBTOP>2.0.CO;2).
- , J. Vialard, A. T. Weaver, D. L. T. Anderson, R. Kleeman, and J. R. Johnson, 2003: The role of air–sea interaction in controlling the optimal perturbations of low-frequency tropical coupled ocean–atmosphere modes. *J. Climate*, **16**, 951–968, [https://doi.org/10.1175/1520-0442\(2003\)016<0951:TROASI>2.0.CO;2](https://doi.org/10.1175/1520-0442(2003)016<0951:TROASI>2.0.CO;2).
- Mu, M., Y. Yu, H. Xu, and T. Gong, 2013: Similarities between optimal precursors for ENSO events and optimally growing initial errors in El Niño predictions. *Theor. Appl. Climatol.*, **115**, 461–469, <https://doi.org/10.1007/s00704-013-0909-x>.
- Osafune, S., T. Doi, S. Masuda, N. Sugiura, and T. Hemmi, 2014: Estimated State of Ocean for Climate Research (ESTOC). JAMSTEC, accessed 25 November 2016, <https://doi.org/10.17596/0000106>.
- , S. Masuda, N. Sugiura, and T. Doi, 2015: Evaluation of the applicability of the Estimated State of the global ocean for Climate Research (ESTOC) data set. *Geophys. Res. Lett.*, **42**, 4903–4911, <https://doi.org/10.1002/2015GL064538>.
- Palmer, M. D., and Coauthors, 2019: Adequacy of the ocean observation system for quantifying regional heat and freshwater storage and change. *Front. Mar. Sci.*, **6**, 416, <https://doi.org/10.3389/fmars.2019.00416>.
- Penduff, T., and Coauthors, 2014: Ensembles of eddy ocean simulations for climate. CLIVAR Exchanges, No. 19, International CLIVAR Project Office, Southampton, United Kingdom, 26–29, https://www.coaps.fsu.edu/pub/eric/papers_html/Chassignet_et_al_14b.pdf.
- Pham, D. T., J. Verron, and L. Gourdeau, 1998: Singular evolutive Kalman filters for data assimilation in oceanography. *C. R. Acad. Sci. Ser. IIA: Earth Planet. Sci.*, **4**, 255–260.
- Qiu, B., D. L. Rudnick, S. Chen, and Y. Kashino, 2013: Quasi-stationary North equatorial undercurrent jets across the tropical North Pacific Ocean. *Geophys. Res. Lett.*, **40**, 2183–2187, <https://doi.org/10.1002/grl.50394>.
- Rhein, M., and Coauthors, 2013: Observations: Ocean. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 255–316, https://www.ipcc.ch/site/assets/uploads/2018/02/WG1AR5_Chapter03_FINAL.pdf.
- Riser, S. C., and Coauthors, 2016: Fifteen years of ocean observations with the global Argo array. *Nat. Climate Change*, **6**, 145–153, <https://doi.org/10.1038/nclimate2872>.
- Roemmich, D., and J. Gilson, 2011: The global ocean imprint of ENSO. *Geophys. Res. Lett.*, **38**, L13606, <https://doi.org/10.1029/2011GL047992>.
- , W. John Gould, and J. Gilson, 2012: 135 years of global ocean warming between the challenger expedition and the Argo Programme. *Nat. Climate Change*, **2**, 425–428, <https://doi.org/10.1038/nclimate1461>.
- , J. Church, J. Gilson, D. Monselesan, P. Sutton, and S. Wijffels, 2015: Unabated planetary warming and its ocean structure since 2006. *Nat. Climate Change*, **5**, 240–245, <https://doi.org/10.1038/nclimate2513>.
- , and Coauthors, 2019: On the future of Argo: A global, full-depth, multi-disciplinary array. *Front. Mar. Sci.*, **6**, 439, <https://doi.org/10.3389/fmars.2019.00439>.
- Sérazin, G., and Coauthors, 2017: A global probabilistic study of the ocean heat content low-frequency variability: Atmospheric forcing versus oceanic chaos. *Geophys. Res. Lett.*, **44**, 5580–5589, <https://doi.org/10.1002/2017GL073026>.
- Sévellec, F., and S. S. Drijfhout, 2018: A novel probabilistic forecast system predicting anomalously warm 2018–2022 reinforcing the long-term global warming trend. *Nat. Commun.*, **9**, 3024, <https://doi.org/10.1038/s41467-018-05442-8>.
- Shea, D., 2013: The climate data guide: Empirical Orthogonal Function (EOF) analysis and rotated EOF analysis. National Center for Atmospheric Research Staff (Eds), accessed 15 November 2022, <https://climatedataguide.ucar.edu/climate-data-tools-and-analysis/empirical-orthogonal-function-eof-analysis-and-rotated-eof-analysis>.
- Stammer, D., 2005: Adjusting internal model errors through ocean state estimation. *J. Phys. Oceanogr.*, **35**, 1143–1153.
- Sun, C., and Coauthors, 2010: The data management system for the global temperature and salinity profile programme. *Proc. OceanObs: Sustained Ocean Observations and Information for Society (Vol. 2)*, Venice, Italy, ESA Publication, WPP-306, <https://doi.org/10.5270/OceanObs09.cwp.86>.
- Tandeo, P., and Coauthors, 2015: Combining analog method and ensemble data assimilation: Application to the Lorenz-63 chaotic system. *Machine Learning and Data Mining Approaches to Climate Science*, Springer, 3–12.
- , P. Ailliot, M. Bocquet, A. Carrassi, T. Miyoshi, M. Pulido, and Y. Zhen, 2020: A review of innovation-based methods to jointly estimate model and observation error covariance matrices in ensemble data assimilation. *Mon. Wea. Rev.*, **148**, 3973–3994.
- Trenberth, K. E., and D. P. Stepaniak, 2001: Indices of El Niño evolution. *J. Climate*, **14**, 1697–1701, [https://doi.org/10.1175/1520-0442\(2001\)014<1697:LIOENO>2.0.CO;2](https://doi.org/10.1175/1520-0442(2001)014<1697:LIOENO>2.0.CO;2).
- von Schuckmann, K., J.-B. Sallée, D. Chambers, P.-Y. Le Traon, C. Cabanes, F. Gaillard, S. Speich, and M. Hamon, 2014: Consistency of the current global ocean observing systems from an Argo perspective. *Ocean Sci.*, **10**, 547–557, <https://doi.org/10.5194/os-10-547-2014>.
- , and Coauthors, 2023: Heat stored in the Earth system 1960–2020: Where does the energy go? *Earth Syst. Sci. Data*, **15**, 1675–1709, <https://doi.org/10.5194/essd-15-1675-2023>.
- Wang, F., J. Feng, Q. Wang, L. Zhang, S. Hu, and D. Hu, 2020: Instability in boundary layer between the North Equatorial Current and underlying zonal jets based on mooring observations. *J. Oceanol. Limnol.*, **38**, 1368–1381, <https://doi.org/10.1007/s00343-020-0015-8>.
- Wilson, C., and V. J. Coles, 2005: Global climatological relationships between satellite biological and physical observations and upper ocean properties. *J. Geophys. Res.*, **110**, C10001, <https://doi.org/10.1029/2004JC002724>.
- Wyrtki, K., 1974: Equatorial Currents in the Pacific 1950 to 1970 and their relations to the trade winds. *J. Phys. Oceanogr.*, **4**, 372–380, [https://doi.org/10.1175/1520-0485\(1974\)004<0372:ECITPT>2.0.CO;2](https://doi.org/10.1175/1520-0485(1974)004<0372:ECITPT>2.0.CO;2).
- Yang, L., R. Murtugudde, S. Zheng, P. Liang, W. Tan, L. Wang, B. Feng, and T. Zhang, 2022: Seasonal variability of the Pacific South Equatorial Current during the Argo era. *J. Phys. Oceanogr.*, **52**, 2289–2304, <https://doi.org/10.1175/JPO-D-21-0311.1>.
- 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. *J. Atmos. Oceanic Technol.*, **37**, 1697–1711, <https://doi.org/10.1175/JTECH-D-20-0001.1>.